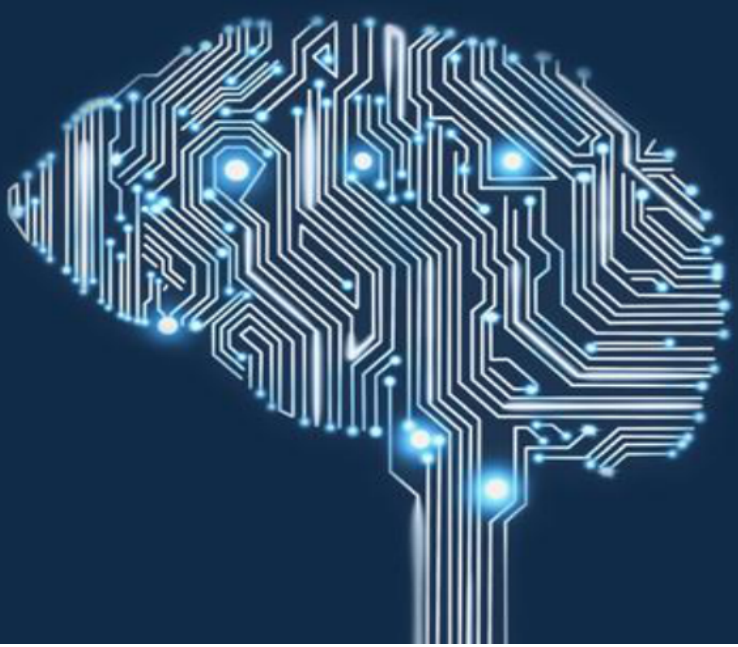
Libros.com

Juan Ignacio Rouyet Ruiz

# ESTUPIDEZ ARTIFICIAL

Cómo usar la inteligencia artificial sin que ella te utilice a ti



Primera edición digital: mayo 2023
Campaña de crowdfunding: equipo de Libros.com
Composición de la cubierta: Mariona Sánchez
Maquetación: Irene E. Jara
Corrección: Beatriz García
Revisión: Isabel Bravo de Soto

**Derechos de reproducción**: Página 18, Group I, *Primordial Chaos*, No. 16, Hilma af Klint de las series WU/Rose 906-1907, colección Courtesy of Stiftelsen Hilma af Klints Verk, imagen digital 2023© Foto Scala Florence/Heritage Images. Página 54, *Mountains and Sea*, Helen Frankenthaler, 1952, detalle de la exposición "Action/Abstraction: Pollack, de Kooning, and American Art, 1940-1976", 4 de mayo - 21 de septiembre 2008, imagen digital 2023© Foto The Jewish Museum/Art Resource/Scala Florence. Página 86, *La traición de las imágenes (Esto no es una pipa)*, René Magritte, 1928-1929, imagen digital 2023© Foto Museum Associates/ LACMA/Art Resource NY/Scala, Florence. Página 116, *Mujer con abanico*, María Blanchard, 1916, imagen digital 2023© Foto Museo Nacional Centro de Arte Reina Sofía. Página 174, *Latas de sopa Campbel*, Andy Warhol, 1962, imagen digital 2023© Foto The Museum of Modern Art, New York/Scala Florence Versión digital realizada por Libros.com © 2023 The Andy Warhol Foundation for the Visual Arts, Inc. / VEGAP
© Helen Frankenthaler, René Magritte, VEGAP, Madrid, 2023

© 2023 Juan Ignacio Rouyet
© 2023 Libros.com

editorial@libros.com

ISBN digital: 978-84-19435-27-9

**Juan Ignacio Rouyet**

# Estupidez artificial

**Cómo usar la inteligencia artificial sin que ella te utilice a ti**

*A Paloma, un potosí.*

# Índice

Portada
Créditos
Título y autor
Dedicatoria

**De agua, harina y sal**

**Miedo me da**

El ferrocarril
El telégrafo
El teléfono
La inteligencia artificial
Autonomía y justicia
Mis conclusiones. ¿Las tuyas?

**Inteligencia probable**

Di lo más habitual
Mira y aprende
Mira y copia
Decide lo más deseado
Somos predecibles
Mis conclusiones. ¿Las tuyas?

**Esto no es una pipa**

¿Hay alguien ahí?
Una cuestión de arte
Una pipa sin consciencia de ser pipa
Mis conclusiones. ¿Las tuyas?

**Será por éticas**

¿Qué es esto de la ética?

Porque buscamos la felicidad
Según cada uno
Lo que nosotros digamos
Un cuadro cubista
Mis conclusiones. ¿Las tuyas?

**Cómo no ser una sopa de datos**

Ética aplicada
Estos son mis principios, pero tengo otros
Seamos éticos, que no cuesta tanto
Evitar la inteligencia artificial pop
Mis conclusiones. ¿Las tuyas?

**Los robots no harán yoga**

**Fuentes, por si quieres seguir bebiendo**

**Agradecimientos**

Mecenas
Contraportada

# De agua, harina y sal

Por el lejano valle de Wakhan, más allá de los montes de Hindu Kush, caminaba el sabio sufí conocido por todos como el Gran Yusuf ibn Tarum. En la cercana comunidad de Ishkashim, se enteraron de la presencia por la zona del gran maestro y salieron por los caminos a su encuentro. Cuando dieron con él, le llamaron y le pidieron que pasara unos días con ellos. El maestro accedió y los acompañó hasta su aldea.

Al día siguiente, toda la comunidad se reunió en la plaza para escuchar las sabias palabras del Gran Yusuf ibn Tarum. El líder de la comunidad se puso en pie para pedir al maestro por sus enseñanzas.

—Gran Yusuf ibn Tarum, sabemos de vuestra grandeza, de vuestra gran sabiduría y que estáis tocado por un espíritu de revelación. Nosotros somos una comunidad pequeña y es posible que nunca hayáis oído hablar de nosotros, de nuestra dedicación y búsqueda de la verdad. Por ello, gran maestro, antes de escuchar vuestras sabias palabras, dejadnos que os contemos nuestro pensamiento y forma de obrar para alcanzar la perfección, de tal forma que podáis confirmar, refutar o completar nuestras ideas.

El Gran Yusuf ibn Tarum se puso en pie e interrumpió al líder, cuando este se disponía a explicar su escuela de pensamiento. Al momento, comenzó a contar las ideas que preocupaban a aquella comunidad, los pensamientos que les hacían errar, las dudas que tenían y cómo intentaban superar sus dificultades. Todos quedaron asombrados de cómo el gran maestro sabía tanto sobre ellos, siendo como eran, una comunidad tan modesta y desconocida.

Cuando terminó, el líder de la comunidad se levantó de nuevo y alabó las palabras del maestro:

—¡Oh, Gran Yusuf ibn Tarum! Es asombroso lo que hemos visto. Se nota con certeza que os acompaña un espíritu de revelación, y así conocéis aquello que a otros se les oculta. Decidnos, gran maestro, cuál es el camino para tal perfección de sabiduría.

El maestro se quedó callado. Se hizo el silencio en la aldea. Tras unos segundos de expectación, pidió que le trajeran un cuenco de barro y agua,

harina y sal. En el cuenco echó el agua, la harina y la sal. Lo removió bien y preguntó al líder.

—Dime, ¿de qué está hecha esta mezcla?

—De agua, harina y sal.

—¿Cómo lo sabes?

—Porque conozco los ingredientes, gran maestro. Si conoces los ingredientes de una mezcla, eres capaz de conocer la naturaleza de la mezcla.

Así ocurre con la naturaleza humana. Conozco vuestros pensamientos porque conozco los ingredientes de la naturaleza humana[1].

Conocer los ingredientes de la mezcla para evitar la estupidez artificial. ¿Qué ingredientes? ¿Qué mezcla? ¿Qué estupidez? Empecemos por eso de la estupidez, que siempre es más simpático.

La inteligencia artificial viene precedida de amenazas, pero también promete esperanzas. Alcanzar las esperanzas o hacer realidad las amenazas es algo que depende de nosotros. De nuestra actuación, es decir, de nuestra inteligencia o de nuestra estupidez.

La estupidez artificial es aquel comportamiento que anula nuestra autonomía cuando usamos la inteligencia artificial. Es cuando decidimos dejar de ser responsables, porque abandonamos nuestra capacidad de responder y dejamos que la inteligencia artificial responda por nosotros. Estupidez artificial es argumentar con frases del estilo «lo dice el sistema», «el algoritmo ha determinado que…», o, la mejor de todas, «la inteligencia artificial ha decidido…». ¡¿Cómo?! ¡Y tú qué! ¿Tú qué dices, qué determinas o qué decides? En todas estas frases falta un «yo» que responde, y en su lugar se traslada la respuesta a un algoritmo.

El primer paso para ser ético no es tanto ser buena persona. El primer paso es tener la voluntad de responder de tus actos ante ti y ante los demás. Delegar la respuesta de tus actos en una inteligencia artificial es abandonar toda responsabilidad. Es dejar de ser ético. Es estupidez artificial. ¿Cómo evitarlo? Conociendo los ingredientes de la mezcla.

La inteligencia artificial es muy compleja de entender. Tanto, que parece magia. ¡Cómo es posible que haga lo que nosotros hacemos! Sin embargo, cuando conoces un truco de magia, carece de emoción. Hubo un tiempo en el que parecía magia que los mensajes fueran instantáneos con el telégrafo, o que el teléfono transportara la voz. También en el pasado tuvimos miedo del tren. Se pensó que nos volvería locos. Hoy estos inventos no nos asombran. Si conocemos los ingredientes, conoceremos la mezcla y lo entendemos todo. Pero en este juego de la inteligencia artificial no hay una mezcla, sino dos: la mezcla de la inteligencia artificial y nuestra mezcla; sí, nosotros mismos.

Te propongo conocer los ingredientes de la inteligencia artificial para quitarle todo atisbo de divinidad. Si queremos usar bien una herramienta tenemos que saber algo sobre cómo funciona. No sabemos construir un coche,

pero sabemos por qué se mueve. No sabemos pilotar un avión, pero sabemos que no vuela por una ciencia desconocida. Hoy vemos que la inteligencia artificial es capaz de entendernos, de hablar, de decidir una ruta, de diagnosticar una enfermedad, de escribir un poema o pintar un cuadro. ¿Magia? No. Nada de eso. Basta conocer sus elementos básicos. Si conocemos los ingredientes de la mezcla de la inteligencia artificial, sabremos usarla, y empezaremos a alejarnos de la estupidez artificial.

También te propongo conocer nuestra mezcla. Saber de nosotros, si nuestros ingredientes son los mismos que los de la inteligencia artificial, y cómo respondemos de lo que hacemos. Este libro tiene algo de filosofía y algo de ética. Quizás pienses que la filosofía no sirve para nada, porque no lleva a conclusiones prácticas. Tienes parte de razón en ello, porque en ocasiones los textos filosóficos no hay quien los entienda. Pero la filosofía te hace pensar, y pensar te ayuda a ser libre. ¿Hay algo más práctico que ser libre?

Te planteo dos objetivos con este libro: conocer y actuar. Conocer los ingredientes para conocer la mezcla. Conocer para actuar, lejos de la estupidez artificial.

Para ello, en cada capítulo de este libro, intentaré ir respondiendo a una serie de preguntas que nos permitirán ir avanzando poco a poco.

¿Debemos tener miedo de la inteligencia artificial?

¿Cuáles son los verdaderos riesgos éticos de la inteligencia artificial?

¿Cómo funciona la inteligencia artificial?

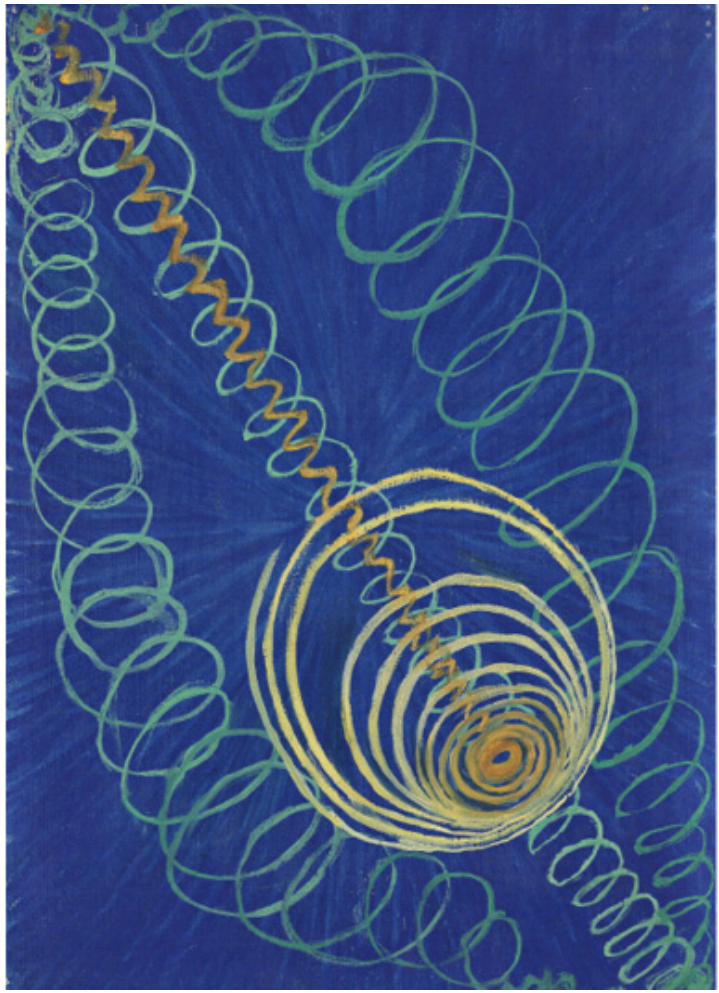
¿Es la inteligencia artificial igual a nosotros?

¿Es posible tener una inteligencia artificial ética?

¿Qué deben hacer las organizaciones para tener una inteligencia artificial ética? ¿Qué debemos hacer nosotros?

Al final de cada capítulo te daré mi respuesta. Pero solo será mi respuesta, no la tuya. Saca tu espíritu crítico y busca tu punto de vista.

En la historia del gran maestro sufí Yusuf ibn Tarum hay dos posibles relatos. En uno de ellos, la inteligencia artificial es quien conoce nuestra mezcla y sabe de nosotros, de cómo actuamos. Estupidez artificial. En el otro escenario, nosotros conocemos la mezcla de la inteligencia artificial y sabemos cómo usarla. Sabiduría natural. Este libro te propone pensar y actuar para hacer realidad este segundo relato.



***Caos primordial*, Nº 16, Hilma af Klint, 1906-07. Fundación Hilma af Klint**

# Miedo me da

Hoy vemos este cuadro y no nos causa rechazo. Nos gustará más o menos, pero lo aceptamos. *Caos primordial* forma parte del llamado arte abstracto, al cual ya nos hemos acostumbrado. Pero en 1906, cuando fue pintado por la artista sueca Hilma af Klint, dentro de una serie de pinturas llamadas *Las pinturas para el templo*, la situación era completamente distinta. Hilma era seguidora del movimiento filosófico-religioso llamado teosofía, que busca el conocimiento de una realidad espiritual que va más allá de las doctrinas particulares de cada religión. Inspirada por esta idea, Hilma pintó su serie de cuadros para el Templo, casi de una forma instintiva, dejando llevar su mano como guiada por una fuerza superior.

En 1908, Hilma enseñó su colección de 111 pinturas abstractas a Rudolph Steiner, filósofo defensor de la teosofía. Rudolph desanimó a Hilma a continuar con aquella línea de pintura, porque era inapropiado para la teosofía. Hilma se quedó desolada y estuvo sin pintar unos 4 años. Posteriormente, continuó con su visión de un nuevo estilo de pintura, dejando una colección de más de 1200 cuadros abstractos.

Cuando murió, en 1944, legó su obra a su sobrino Erik, y dejó escrito en su testamento que su obra no se hiciera pública hasta 20 años después de su muerte, esperando que, para entonces la sociedad, en general, pudiera entender su arte. En 1970, cumplido el plazo de los 20 años, Erik presentó las obras al Moderna Museet de Estocolmo, quien las rechazó cuando se enteró que la autora había flirteado con la teosofía. La primera exposición pública de la obra de Hilma af Klint no fue hasta 1986, en Los Ángeles, bajo el título «Lo espiritual en el Arte: pinturas abstractas 1890 – 1985»[2].

Todo este desprecio por la obra de Hilma af Klint, por su relación con una visión espiritual, dio fama a Vasili Kandinsky, quien es considerado como el padre de la pintura abstracta. En 1911 publicó su famosa obra *De lo espiritual en el arte*[3], donde sienta las bases del arte abstracto y explica cómo una serie de formas y colores pueden inspirar ciertas ideas y emociones. A partir de entonces, sus obras tituladas como *Composiciones* (para qué darle nombres

concretos) forman parte de la historia del arte contemporáneo y del arte abstracto.

En 1906 el arte abstracto de Hilma fue rechazado, pero 5 años después comenzó a cambiar la visión sobre el arte abstracto. En los años siguientes se aceptaron nuevas normas y se hicieron nuevos juicios de valor. La desconocida Hilma pasó 80 años en el olvido hasta que comenzó a ser reconocida su aportación al arte. Ahora sus pinturas no nos sorprenden, ni nos causa rechazo si han sido inspiradas por una visión espiritual.

En esta historia hay un juicio de valor erróneo y un acierto. El juicio de valor erróneo consistió en denostar una obra de arte por las inclinaciones filosóficas de su autora. El error de creer que una obra de arte debería estar realizada por un artista con unas condiciones determinadas, lo que sería considerado un artista serio. «Debería», ya ha salido la palabra. El acierto estuvo en Kandinsky, quien antes de exponer una obra abstracta, explicó en un libro cómo entenderla.

Lo ocurrido con la obra de Hilma af Klint ocurre ahora con la inteligencia artificial. En cualquier actividad que realizamos, siempre existe un debate entre lo que es y lo que debería ser. La obra *Caos primordial* no era como debía ser. Ese tipo de arte era entonces inaceptable, pero hoy en día ya es aceptable. Incluso aceptamos cualquier tipo de expresión artística. No hay más que pasarse por ARCO. En el arte ya no hay deberías. Y la inteligencia artificial, ¿es cómo debería ser?, ¿dejará de tener deberías?, ¿debemos tener miedo a la inteligencia artificial?

Para responder a estas preguntas podemos ver lo que ha ocurrido en el pasado con otras tecnologías. Te propongo un viaje por tres avances tecnológicos del siglo XIX: el ferrocarril, el telégrafo y el teléfono. En los tres casos vamos a ver el debate que hubo en su momento sobre lo que debería y no debería ser, y cómo este debate ha ido cambiando con el tiempo. Veremos que los miedos iniciales —lo que no debería ser— no eran tan peligrosos, y que muchas cuestiones que eran inaceptables en aquella época, hoy las tenemos por cosas normales. Igualmente, había muchas esperanzas —lo que sí debería ser—, que con el tiempo se han visto que no han llegado a ser realidad. Lo que ha ocurrido con otras tecnologías, ocurrirá con la inteligencia artificial.

Al juzgar el arte de Hilma af Klint los deberías se centraron en si su obra seguía un estándar artístico o si su biografía era seria. Si hablamos de inteligencia artificial, ¿sobre qué habla lo que debería o no debería ser?

Actualmente en nuestra sociedad todo debate se centra en cuatro dimensiones que articulan lo que debería y no debería ser. Son las dimensiones de salud, seguridad, economía y moral. No necesariamente por este orden, aunque las cuestiones morales suelen surgir en último lugar. Tanto defensores como detractores crean sus argumentos en base a todas o algunas de estas dimensiones. Juicios de valor que, al igual que en el arte, han ido cambiando, y

lo que antes era «esta tecnología no debería ser así», con el tiempo se ha convertido en «sí debe ser así». Vamos a verlo.

## El ferrocarril

### *Velocidad que enloquece*

La irrupción del ferrocarril a comienzos del siglo XIX unió los pueblos y abrió los miedos. En la eterna, tranquila y verde campiña británica apareció un monstruo: la máquina de vapor. El ferrocarril fue la imagen viva del progreso tecnológico y su presencia dividió tanto a los campos, por el tendido de raíles; como a la sociedad, por los «deberías».

La primera cuestión sobre lo que debería o no debería ser vino con la seguridad, y, en particular, de la mano de la velocidad. Viajar a 50 Km/h se percibía como un riesgo considerable. Este miedo se vio acrecentado por dos accidentes memorables en sus comienzos.

Uno de ellos sucedió el mismo día de la inauguración de la línea Liverpool-Mánchester en 1830. Durante una parada del tren para abastecerse de agua, William Huskisson, miembro del Parlamento, fue atropellado por una locomotora Rocket debido una imprudencia. Murió a los pocos días[4]. Años más tarde, el escritor Charles Dickens tuvo un accidente de tren. Dos raíles sobre un viaducto fueron retirados por error al mirar un horario de tren equivocado. Todos los vagones de primera acabaron sobre el río, excepto el de Dickens, que quedó colgando[5]. Dickens se salvó, pero no así la reputación del ferrocarril como transporte seguro.

Aquí tenemos ya un paralelismo con la inteligencia artificial y los vehículos de conducción autónoma. Existen dudas sobre su seguridad. No ayuda el accidente de 2018 en Arizona, donde un vehículo autónomo de Uber mató a una mujer, que cruzaba la carretera con su bicicleta. Tuvo repercusión por dos motivos: fue el primer accidente por atropello de un vehículo autónomo y fue causado por un error de *software*. El radar no identificó correctamente a la señora cruzando la carretera, y además Uber había deshabilitado la opción de frenado de emergencia[6]. En el accidente de Dickens se quitó un rail por error; en el caso de Uber se deshabilitó una opción del *software*.

Todo avance tecnológico tiene un periodo inicial de desgracias. Dicho de una manera quizás más radical: todo avance tecnológico ofrece nuevas formas de morir; no se sabe de nadie que muriera por accidente de ferrocarril en la época del Imperio romano. A pesar de estos siniestros, el ferrocarril ha continuado y hoy es un medio de transporte bastante seguro. La inteligencia artificial también llegará a ser segura.

No obstante, no era necesario sufrir un accidente para padecer daños en la salud. Se pensaba que la velocidad en sí misma causaba demencia. Así se atestiguaba en múltiples casos de personas, particularmente hombres, que durante un viaje en tren habían tenido comportamientos violentos y actitudes fuera de sí[7]. Cada vez que el tren paraba en una estación, estas personas

recobraban su sensatez y se comportaban de manera correcta y educada. No obstante, una vez que el tren se ponía en movimiento, la actitud violenta y errática volvía a manifestarse.

Se pensaba que el traqueteo del tren, unido al excesivo ruido por el movimiento sobre los raíles, literalmente afectaba al cerebro y alteraba los nervios. Se culpó, entonces, no solo a la máquina de vapor como el instrumento de progreso que no debía existir, sino que también se culpó a la propia civilización como el origen de nuestros males. En la última mitad del siglo XIX, los exploradores que volvían de las zonas más remotas del mundo informaban que apenas veían personas con enfermedades mentales en las áreas con menor civilización. Se llegó entonces a establecer que la locura, o cualquier otra enfermedad mental, era la inevitable consecuencia de la civilización y que un incremento de la demencia era el castigo que debíamos pagar por el incremento de la civilización[8].

¿La civilización afecta a nuestro comportamiento? Todo parece apuntar a que sí. De hecho, ese es el objetivo de la civilización: tener un comportamiento específico, que denominamos *social*. ¿Nos lleva a comportamientos no deseados, como la locura con el tren? Esto es más discutible. En el caso del ferrocarril con el tiempo se ha visto que no. Con la inteligencia artificial, se empezó a analizar si los asistentes de voz, tipo Alexa (Amazon), Google Home, Siri (Apple) o Cortana (Microsoft), afectaban al comportamiento de los niños. Habitualmente, cuando hablamos con estos dispositivos, no decimos palabras tales como «por favor» o «gracias», más bien decimos simplemente: «Alexa, ¿qué hora es?». Se vio que los niños usaban este estilo directo de comunicación, tanto con los asistentes de voz, como en la comunicación con los adultos. Parecía que estos siervos inteligentes volvían maleducados a los niños.

Esta situación llevó a Amazon a incorporar la «Magic Word» en su sistema de Alexa. Con esta nueva funcionalidad, cada vez que se pide algo a Alexa incluyendo la palabra mágica «por favor», el sistema responde también de manera agradecida, diciendo, por ejemplo, «gracias por ser tan amable». De manera similar, Google integró posteriormente su funcionalidad «Pretty Please».

Posteriormente se vio que estos asistentes inteligentes no afectaban al comportamiento de los adultos y que los niños eran conscientes de cuándo hablaban con una máquina o con una persona. Al final todo se aclara y nosotros nos adaptamos.

Ahora debemos dar paso a los economistas. En aquel entonces del ferrocarril, todo parecía indicar que la civilización se veía atacada por la innovación. Tocaba, entonces, hablar de los importantes beneficios económicos que traería la innovación del ferrocarril.

### *Los zapateros venderán más zapatos*

Claramente, el ferrocarril era más productivo y eficiente, comparado con los medios habituales de transporte del momento, que eran a caballo o en barcazas por canales fluviales. Las palabras *productivo* y *eficiente* nunca deben faltar si queremos demostrar la bondad de una innovación.

El ferrocarril iba a generar un gran beneficio a toda la sociedad. Para convencer de ello, lo mejor era demostrarlo con un ejemplo cercano, como era el caso de suponer la existencia de un humilde zapatero en un remoto pueblo campestre. Gracias al tren, este zapatero vería ampliado su mercado, porque ya no solo vendería calzado a sus vecinos del pueblo, sino que podría vender sus zapatos en las principales ciudades británicas[9]. El ferrocarril era una fuente de negocio para los zapateros y para cualquier empresario o dueño de un humilde negocio.

Pero no solo para dueños de negocio. El ferrocarril iba a permitir ahorrar dinero y tiempo y tener una vida más plena. Cualquier mercancía transportada por canal entre Mánchester y Liverpool llevaba 36 horas de viaje, frente a la hora y tres cuartos que suponía transportarla por tren, lo que supone un ahorro del 50 % del coste. Tomando como base estas eficiencias, vinieron los análisis de grandes números con grandes esperanzas.

Teniendo en cuenta que medio millón de personas al año utilizaban dicha línea de tren, si cada uno de ellos ahorraba tan solo una hora en el trayecto, esto suponía un ahorro de 500.000 horas, es decir, 50.000 días de trabajo —en aquel entonces una jornada de trabajo eran 10 horas al día—. Esto equivalía a aumentar la fuerza de trabajo en 160.000 hombres, sin necesidad de incrementar la comida para alimentarlos. Hemos conseguido más capacidad de trabajo. Además, el trabajo de estos 160.000 hombres sería de más valor que el de aquellos ocupados en el campo o en el transporte convencional a caballo o barcazas[10].

Es posible que todo esto nos suene. Toda innovación siempre nos promete más valor añadido, lo que quiera que signifiquen las palabras *valor* y *añadido*. Nunca nos lo explican, tan solo sueltan el par de palabras acompañadas de una sonrisa de satisfacción. Está claro que, si se quiere vender la bondad de una tecnología, hay que meter la palabra valor. La inteligencia artificial también nos promete trabajos de más valor. Aquellas actividades tediosas, rutinarias o peligrosas las podrá hacer un robot que ni siente ni padece. El tema merece más atención y lo veremos posteriormente, pues, por otro lado, si la inteligencia artificial hace un trabajo tedioso, quiere decir que una persona deja de tener un trabajo. Aunque fuera tedioso, seguro que le arreglaba la vida. Lo mismo sucedía con el ferrocarril, su eficiencia venía a cambio de una pérdida.

El ferrocarril iba a eliminar puestos de trabajo: aquellos dedicados, precisamente, al transporte por caballo o canales, lo cual generaría pobreza en las personas dedicadas a tal actividad[11]. Sin embargo, un avispado cochero no tendría por qué preocuparse.

Se estaba demostrando que el transporte a caballo, debido a la línea férrea entre Mánchester y Liverpool, iba en aumento al favorecer el transporte de corta distancia[12]. Las grandes distancias eran cubiertas por ferrocarril, pero esos tramos cortos entre una estación y una población cercana eran salvados por diligentes taxistas a caballo. Como el tren movía a muchas a personas, estos audaces conductores verían incrementado su negocio.

Todo eran parabienes. Algunos oficios, como el transporte por canal, se verían afectados por la llegada del ferrocarril, pero el resultado global sería que los zapateros venderían más zapatos, los cocheros llevarían a más gente, el país sería más productivo y habría trabajo de más valor. El tren sería más productivo, más eficiente y de más valor: ¿cómo negarse ante esas tres palabras?

La inteligencia artificial quizás nos quitará puestos de trabajo, pero nos dicen que haremos otros y de más valor. Al igual que los cocheros de entonces, nos dicen que no nos preocupemos. Más adelante veremos si debemos preocuparnos o no. De momento, el tren todavía traerá más enhorabuenas.

### *Peor lana, pero más inteligentes*

Finalmente, quedaron los argumentos de naturaleza moral, que eran aquellos que asignaban un cierto juicio de valor al invento. El ferrocarril fue considerado una máquina infernal. El propio nombre representaba un cierto aire de odio: ferrocarril, carril de hierro; hierro, metal, que denota dureza y batalla. ¡Qué cosa tan inhumana!

El tren se consideraba innecesario, pues no había razón de viajar tan rápido, y además iba en contra de los valores de tranquilidad, belleza y unidad. La principal oposición vino por los dueños de las tierras, que veían cómo aquella máquina oscura y agresiva atravesaba sus campos expulsando humo negro a toda velocidad.

Aquello no podía traer nada bueno: rompía la paz de un desayuno tranquilo y llenaba sus hogares de hollín y ruido; destruía su privacidad, pues los pasajeros de los trenes pasaban cerca de sus casas; destruía la unidad de las granjas, al ser divididas por los raíles del tren; desfiguraba el paisaje, al cortar montañas o crear largos puentes para cruzar valles o ríos; interfería en la caza; e incluso, y aquí empieza lo verdaderamente pintoresco, afectaba a la calidad de la lana de las ovejas que pacían junto a las vías del tren. En resumen, el ferrocarril transmitía los valores de vulgar, mercenario y desagradable[13]. Claramente, el ferrocarril no era lo que debía ser, pues no traía nada bueno.

Sin embargo, frente a esta visión tan deprimente, existía otra forma de ver las cosas. También había grandes esperanzas para la humanidad en aquellas máquinas de hierro y carbón. Dado que el ferrocarril iba a aumentar la productividad de cada artesano, —por ejemplo, de los zapateros— se iba a producir una cadena de consecuencias virtuosas sin precedentes: al tener que fabricar más zapatos, sería más competente en su trabajo; esto le haría ser más

rápido en hacer zapatos, lo que le daría más tiempo libre; el tiempo libre le induciría a la indagación; y la indagación, al conocimiento[14]. Por tanto, el tren iba a hacer que la gente del siglo XIX fuera más inteligente, al menos, los zapateros.

Pero aún hay más. No era verdad que el ferrocarril fuera a estropear el eterno candor de suaves valles y colinas. Antes bien, la construcción de puentes, viaductos, accesos y depósitos relacionados con el ferrocarril, traerían un nuevo tipo de arquitectura creando un embellecimiento arquitectónico. Esta nueva belleza sería fuente de una mejora en los hábitos y la moral de la población rural: fomentaría el cultivo del gusto y la difusión del conocimiento, por medio de la comunicación global[15]. Se llegó a decir, textualmente y por personas muy sesudas que «la duración de nuestras vidas, en lo que respecta al poder de adquirir información y diseminar conocimiento, se duplicará; y podemos estar justificados al buscar la llegada de un tiempo, cuando el mundo entero se habrá convertido en una gran familia, hablando un solo idioma, gobernado en unidad y armonía por leyes similares, y adorando a un solo Dios»[16].

Con estas palabras, uno queda rendido ante el ferrocarril. Esos trenes que cruzan los campos iban a conseguir que los pueblos fueran más inteligentes, con mejores hábitos, más éticos y más fraternales. Hoy en día, con la alta velocidad, cabe suponer que nuestra inteligencia, nuestra ética y nuestra fraternidad vuelen a la velocidad del rayo. ¿Somos hoy todo eso? ¿Se cumplen las bondades que nos venden con las innovaciones? ¿La felicidad que nos pintan, con solícitos robots a nuestro lado, será verdad?

Hablando de la velocidad del rayo, unos años más tarde, llegó la comunicación a la velocidad de la luz.

## El telégrafo

### *Acabaremos con la barbarie*

No todo avance tecnológico fue tachado, de primeras, de incorrecto. El establecimiento de la primera línea de telégrafo en 1844 trajo grandes esperanzas para la humanidad, casi como el tren[17]. Con el telégrafo, por primera vez, la transmisión de la información se separaba del medio físico por el cual viajaba. Hasta la fecha, la información era transportada por un mensajero, habitualmente a caballo, y esta viajaba a la velocidad a la que iba el mensajero. Si el mensajero tardaba tres días en llegar a su destino, el mensaje tardaba tres días en llegar a su destinatario. Si el mensajero corría, el mensaje llegaba antes; si se entretenía en cantinas por los caminos, la misiva llegaba tardía.

El telégrafo hizo que la información viajara a la velocidad del rayo. Esto era increíble. ¿Cómo era posible que un mensaje llegara de manera instantánea, sin mediar un mensajeo por medio? Además, ¿qué es eso de la electricidad? Para mayor conmoción el telégrafo estaba basado en esa fuerza invisible llamada electricidad, que no todo el mundo llegaba a entender. Parecía mentira que el pensamiento pudiera ir más deprisa que la materia. Si nuestras ideas podían volar a la velocidad de un suspiro, entonces se aventuraban grandes esperanzas.

Se creó el concepto de «comunicación universal» como aquel medio por el cual se podría unir la mente de todos los hombres en una especie de conciencia común. El telégrafo iba a unir a todos los hombres y mujeres de la Tierra para transmitir los principios más elevados del humanismo. Por fin la barbarie estaba llamada a su fin. Era imposible que los viejos prejuicios y las hostilidades existieran por más tiempo, dado que se había creado un instrumento que permitía llevar el pensamiento a cualquier lugar del mundo, y con independencia del mensajero. Ahora sí que estábamos delante de un invento moralmente bueno en sí mismo. El telégrafo era lo que debía ser.

Todas sus supuestas bondades innatas aparecieron reflejadas en la primera frase que se transmitió por dicho medio, atribuyendo al invento una dote divina. El 24 de mayo de 1844 se produjo la primera transmisión pública por telégrafo entre Washington y Baltimore, cubriendo una distancia de unos 60 Km. Samuel Morse, desde la Cámara de la Corte Suprema en el Capitolio de EE. UU., envió a su colega Vail, en Baltimore, la frase «What hath God wrought!», tomada de la Biblia. En particular del capítulo 23 y versículo 23 del Libro de los Números, y que se puede traducir como «¡Lo que Dios ha hecho!». De alguna forma, el telégrafo era una creación divina que iba a traer toda clase de beneficios a la sociedad y era propio de una obra celestial.

Si ya con el ferrocarril íbamos a ser más éticos y fraternales, y ahora con el telégrafo se iba a acabar la barbarie, hoy en día el mundo debe ser maravilloso,

aunque quizás no nos demos cuenta. Más bien, parece que hemos conseguido una comunicación global, enganchados al móvil, pero no tanto esa «comunicación universal» que nos lleva a una conciencia común y a una unidad entre los seres humanos. No hay más que ver ciertos mensajes en las redes sociales.

Hoy vemos que, finalmente, el telégrafo no ha terminado con la barbarie. ¡Qué cosa más rara! ¿Qué puede haber fallado, si todo apuntaba tan bien?

### *La guerra por un telegrama*

El error —la falacia— radicó en creer que una tecnología tiene connotaciones morales y que es buena o mala en sí misma, sin tener en cuenta el uso que nosotros hacemos de ella. El telégrafo no es ni bueno ni malo, depende del uso que hagamos de él, es decir, de los mensajes que transmitamos. Un ejemplo claro es lo que ocurrió con el llamado telegrama de Ems que causó la guerra Franco-Prusiana en 1870. Justo lo contrario de acabar con la barbarie.

En aquellas fechas del siglo XIX, España, una vez más, era el tablero de ajedrez de las vanidades de Europa por cuestión de la sucesión al trono real. Con el exilio de Isabel II tras la Revolución de 1868, conocida como la Gloriosa, se comenzó a buscar un monarca para España. Una de las opciones fue el príncipe Leopoldo de Hohenzollern, propuesto por Otto von Bismarck, en aquel entonces, primer ministro de Prusia. Esto no gustó a Francia, porque suponía aumentar el poder de Prusia en Europa. Finalmente, Francia consiguió que Prusia abandonara su propuesta. Pero Napoleón III, rey de Francia, quiso tener por escrito la renuncia de Guillermo I, rey de Prusia, a toda pretensión de proponer un candidato para el trono español.

Aprovechando que Guillermo I pasaba unos días en el balneario de Ems, Napoleón III envió a un embajador a entrevistarse con él. El rey Guillermo I le recibió, si bien rehusó dejar por escrito dicha renuncia, aduciendo que, por su parte, no tenía noticias oficiales de la renuncia de Leopoldo de Hohenzollern, como, en efecto, así era. Al día siguiente, supo de tal renuncia y, en lugar de entrevistarse de nuevo con el embajador, le mandó un mensaje diciendo que, dado que ya no había candidato, por su parte no había más que decir. Hasta aquí, todo correcto, formal y diplomático. Sin embargo, los intereses particulares pueden trastocar la realidad.

Con la nueva situación, Guillermo I mandó un telegrama a Bismarck relatando los acontecimientos y dejando en manos de este si comunicar o no tal hecho a la prensa y el cómo hacerlo. Bismark era defensor de la candidatura de Leopoldo de Hohenzollern y no le gustó cómo se habían sucedido los acontecimientos. Esto no podía quedar así. Redactó un nuevo telegrama para la prensa alterando ligeramente los hechos. En su telegrama se decía escuetamente que Guillermo I rechazó recibir al embajador y que no tenía nada que decirle. Rechazar recibir a un embajador y proclamar que no hay

nada que decir es una afrenta diplomática. Este rechazo y silencio, que no fueron tales, se entendió como un desprecio a Francia. El telegrama fue enviado el 13 de septiembre de 1870 y Napoleón III declaró la guerra a Prusia 6 días después.

¿Esto es lo que se dice acabar con la barbarie? Esta historia suscitó un debate sobre si la tecnología afectaba a nuestra capacidad de pensar de una forma sosegada. El debate sigue activo hoy, con unas redes sociales que no descansan, movidas por una inteligencia artificial. Veamos primero el fundamento.

### *Sin capacidad de pensar*

Este hecho de la guerra Franco-Prusiana trajo una consideración sobre el telégrafo: no dejaba tiempo para pensar. La inmediatez del telégrafo transmitía el valor de la necesidad de inmediatez en la respuesta. Todo, hasta las respuestas, tenía que suceder a la velocidad de la luz, porque la información iba a la velocidad de la luz. Incluso las noticias se sucedían a la velocidad de la luz.

La noticia del telegrama de Bismark fue publicada y discutida de manera vertiginosa por los periódicos, lo que favoreció un clima social que exigía una respuesta rápida ante tamaña ofensa. Por ello, se culpó al telégrafo en sí por el estallido de la guerra Franco-Prusiana de 1870, no tanto por el contenido del telegrama, sino porque esa comunicación instantánea obligó a imprimir un ritmo acelerado en la diplomacia que impidió la reflexión sosegada de los hechos. En cierta manera, la declaración de guerra seis días después de recibir el telegrama fue el producto de una toma de decisión poco meditada, acuciada por la necesidad de dar una respuesta tan rápida como la rapidez del telégrafo.

El periódico inglés *Spectator* publicó en 1889 el editorial «Los efectos intelectuales de la electricidad» donde alertaba de los daños que podía causar el telégrafo en nuestra mente. En realidad, como se ve por el título del editorial, la preocupación no nacía del telégrafo en sí mismo, sino de la electricidad, que era esa fuerza misteriosa del telégrafo.

Lo más alarmante para el *Spectator* era lo que denominaba «fuerza intelectual» de la electricidad, la cual afectaba al telégrafo. Según el periódico, el telégrafo era un invento que no debía ser, pues su uso iba a afectar al cerebro y al comportamiento humano. Merece la pena leer algunos párrafos, no tienen desperdicio[18]:

> [Debido al telégrafo] todos los hombres están forzados a pensar en todas las cosas al mismo tiempo, en base a información imperfecta y con muy poco tiempo para la reflexión. Es rumor, más que inteligencia, lo que se apresura sin aliento por mares y continentes. […] La constante difusión de declaraciones en fragmentos, la constante emoción de sentimientos no justificados por hechos reales, la constante formación de opiniones apresuradas o erróneas, al final, se podría pensar, acabará por deteriorar la inteligencia de todos aquellos a los que apelan al telégrafo. […] El resultado es una precipitación universal y confusión de juicio, una disposición a

decidir demasiado rápidamente, una impaciencia si no se toman medidas apresuradas antes de que los estadistas u otros responsables hayan tenido tiempo de pensar. Es como si todos los hombres tuvieran que estudiar todas las cuestiones bajo la emoción de la ira, el miedo o la pena, o con esa sensación consciente de estar «agitado» que, de todas las molestias recurrentes de la vida, hace que una verdadera reflexión sea lo más difícil o imposible. […] Esta excitación antinatural, esta perpetua disipación de la mente, esta pérdida de sentimiento en las escenas de una ópera, al final tiene que acabar por dañar la consciencia y la inteligencia; y esto, que se lo debemos a la electricidad, deber ser balanceado respecto a cualquier beneficio material que pueda ofrecer. No decimos nada más que el resultado universal del uso del telégrafo es sobrecargar la mente ordinaria con fragmentos de información indigeridos e indigestibles, e insistimos solo que su tendencia será debilitar y finalmente paralizar el poder reflexivo.

Puedes sustituir la palabra telégrafo por redes sociales, o por cualquiera de sus hijos (Whatsapp, Twiter, Instagram, Facebook, TikTok, YouTube, por citar algunos) y la reflexión parece que siga siendo válida. Esta historia del telégrafo nos muestra cómo una tecnología no es en sí misma ni buena ni mala. Pero tampoco es neutral, porque tiene sesgos a través de los valores que transmite. Por ello, tenemos que tomar una doble decisión cuando estamos con una tecnología: sobre sus fines y sobres sus valores. Determinar para qué queremos la tecnología (para hacer la guerra, para comunicar la paz) y qué valores aceptamos (rapidez o sosiego).

En lugar de la palabra telégrafo puedes poner cualquier otra tecnología. El resultado es el mismo. La diferencia con las tecnologías que tenemos actualmente es que sus fines o sus valores se ven aumentados por la inteligencia artificial. Pero esto lo veremos posteriormente, porque todavía tenemos que dar la bienvenida al teléfono. ¡El que faltaba!

## El teléfono

### *Un invento indigno*

La invención del teléfono hacia mediados del siglo XIX estuvo rodeada de asombro, miedo y suspicacia. Inventado por Antonio Meucci en 1854, fue formalmente patentado por Graham Bell en 1876, quien lo presentó públicamente en la exposición del Centenario en Filadelfia ese mismo año. Hoy en día, en una sociedad que parece móvil-dependiente, apenas podríamos vivir sin nuestro teléfono, pero en aquella primera presentación el invento pasó inadvertido para público y jueces. Tuvo que pasar por el estand de Bell el emperador de Brasil, Pedro II, conocido del propio Bell, quien lo probó y dijo con sorpresa: «¡Dios, esto habla!»[19].

A partir de entonces, el teléfono empezó a ser notorio, pero no apreciado. Se pensó que el teléfono era indigno para las personas. Todo aquel que probaba el invento se sentía estúpido hablando delante de un disco de metal, especialmente cuando tenían que gritar para hacerse escuchar. El teléfono podía ser un gran invento, pero ello no compensaba la pérdida de dignidad personal que suponía su uso.

Entramos de lleno en las consideraciones de valor, en las cuestiones morales. Hablar por teléfono no era moralmente adecuado porque era indigno. Apareció la cuestión de la tecnología y dignidad humana en su uso. El teléfono era una tecnología que nos degradaba como seres humanos. Hoy puede que no pensemos así y no nos causa rubor hablar por el móvil en la calle como si estuviéramos comiendo una rebanada de pan y a viva voz. Ahora pensamos que la inteligencia artificial nos degrada como seres humanos, porque parece que un conjunto de cables es como nosotros y nos supera. Con el tiempo, quizás, acabemos perdiéndole el respeto a la inteligencia artificial, igual que con el teléfono.

En aquel entonces, una de las causas de considerar al teléfono indigno era el desconocimiento de la tecnología que se estaba usando. No se entendía cómo hablando delante de un disco de metal se podía transmitir la voz por un cable. Bell fue acusado de impostor, y se dijo que era un simple ventrílocuo. El teléfono era un engaño y solo los incautos podían pensar que aquello era verdad.

*The London Times* defendió con orgullo que era imposible transmitir la voz por un cable debido a la naturaleza intermitente de la corriente eléctrica. El físico Joseph Henry, famoso por sus estudios en electromagnetismo, aseguró que ese invento era imposible porque contradecía la ley de la conservación de la energía —esa que viene a decir que la energía ni se crea ni se destruye, solo se transforma; ¡a saber por qué pensó eso!—. *The New York Herald* dijo que el efecto de aquel aparato era sobrecogedor y casi sobrenatural y *The Providence*

*Press* dijo que las fuerzas de la oscuridad de alguna forma estaban aliadas con el aparato. Tan solo un mecánico de Boston llegó a dar con la explicación sobre el funcionamiento de aquel extraño aparato parlante. Su funcionamiento se basaba en un orificio a lo largo de todo el cable por el cual se transmitía la voz. Siempre han existido respuestas simplistas que buscan tranquilizar los pensamientos.

El desconocimiento de cómo funciona una tecnología no ayuda a que esta sea aceptada. El teléfono era un engaño, o bien un invento del lado oscuro, porque se ignoraba su funcionamiento. La inteligencia artificial nos asusta porque nos asombra y no comprendemos cómo es capaz de hacer todo lo que vemos de ella. Pensamos que usurpa algunas de nuestras funciones como seres humanos, en particular la capacidad de pensamiento. En capítulos posteriores empezaremos a desmitificar tal asombro.

Si desde un punto de vista moral el teléfono era impropio, desde un punto de vista económico no corría mejor suerte.

### *Ni a banqueros ni a tenderos*

Tras el recelo y la desconfianza vino la indiferencia y el menosprecio. No se veía la utilidad del teléfono. Debemos entender que no se veía su beneficio económico. No había respuesta clara a la pregunta «¿Cómo puedo ganar dinero con este invento?». Se pensaba que estaba bien como divertimento para los científicos de laboratorio, que podían jugar con él para entender el funcionamiento de la electricidad, pero no se apreciaba un uso razonable para el resto de personas. Al menos un uso por el cual se pudiera cobrar.

De forma graciosa se decía que los banqueros pensaban que el teléfono podría ser útil para los tenderos, si bien nunca sería útil para los banqueros; y los tenderos decían que podría ser útil para los banqueros, pero nunca útil para los tenderos. Ninguno de los dos vio negocio en el teléfono. ¿Qué pasaría si les preguntáramos hoy?

*The Boston Times*, en un editorial burlón, dijo que con el teléfono uno podía cortejar a una chica, ya estuviera en China o en Boston. Pero el aspecto más serio y alarmante del invento, continuaba el editorial de forma irónica, era el poder irresponsable que daría a la mayoría de las suegras, las cuales serían capaces de enviar su voz por todo el globo habitable[20]. En resumen, no se veía utilidad seria —es decir, económica— al teléfono y solo era la diana apropiada para dardos satíricos.

Una vez que hemos desprestigiado al teléfono, corresponde desprestigiar a su inventor. Si la pobre Hilma af Klint fue ninguneada después de su muerte por sus inclinaciones espirituales, Graham Bell no corrió mejor suerte, porque no se veía la necesidad de su invento. Cuando pensamos que una obra no es como nosotros pensamos que debe ser, pasamos a desacreditar al autor de la obra. ¡Fuera Hilma, fuera Graham Bell! Es la llamada falacia *ad hominem*, que dicho

de manera común significa que, como, en realidad, no tengo argumentos racionales en contra de tus ideas, pues me meto contigo. Es más eficaz porque exige pensar menos. Esta falacia se ve mucho en la esfera pública, entre aquellos que nos gobiernan y otros desgobernados que habitan en las redes sociales.

El teléfono en sus orígenes era un desperdicio tecnológico: ni moralmente aceptable ni económicamente viable. Ahora nos falta abordar la cuestión de la seguridad.

### *Un montón de gelatina transparente*

Según el teléfono se iba implantando en la sociedad, y, en particular, en los hogares, apareció la sombra de la pérdida de privacidad. ¡La seguridad sale a escena! Se percibía como un gran peligro que miles y miles de hogares estuvieran conectados por cables, capaces convertir las intimidades caseras en algo público. Se pensó que la intimidad iba a desaparecer por completo. En las editoriales de los periódicos se podían leer frases tan amenazantes como «¿Qué será de la privacidad de la vida?», o bien, «¿Qué será de la santidad del hogar doméstico?»[21]. La grandilocuencia siempre abona los miedos.

A propósito de la instalación del tendido de líneas de teléfono en el estado de Rhode Island, *The New York Times* publicó que los técnicos que instalaron los cables pudieron escuchar a clérigos parlanchines, canciones melodiosas o gatos de medianoche, así como las conversaciones confidenciales de cientos de matrimonios[22]. La noticia, fuera verdadera o no, revela la preocupación del momento por la privacidad del teléfono. ¡Si vieran la situación hoy en día…!

Partiendo de esta premisa de ataque a la privacidad, el argumento derivó, de nuevo, en aspectos morales. El teléfono iba a afectar a la esencia humana. Nada más y nada menos. El teléfono nos iba a transformar como sociedad y como personas, por esa virtud que tiene de transmitir el pensamiento a distancia. Con el teléfono se rompería la esfera de lo privado, y eso acabaría con lo más profundo del ser humano.

Siempre nos hemos preguntado por la esencia del ser humano. Por aquello que nos distingue de cualquier otro ser vivo, en particular de los animales. En aquella época del teléfono naciente se pensaba que un elemento de la esencia del ser humano era la privacidad. Esta visión tiene su fundamento, pues se desconoce que, por ejemplo, las vacas tengan vida privada —aunque quizás la tengan y esta sea tan privada que no la conocemos—. La vida privada corresponde a esa parte de nuestra existencia sobre la cual no queremos dar cuenta públicamente y que nos identifica como individuos en lo más profundo.

Una publicación británica especializada en tecnología vaticinaba que pronto seríamos el uno para el otro un montón transparente de gelatina[23]. El teléfono nos iba a convertir en una masa social uniforme sin la singularidad propia de cada persona. La frase me parece significativa, porque habla de *transparencia* y de *gelatina*. No sé si por primera vez, pero quizá sí una de las

primeras veces, se hablaba de la transparencia que causa la tecnología y de la pérdida de valor único personal que esto puede suponer. Esa transparencia nos lleva a una masa informe de gelatina, fácilmente moldeable, donde todos somos lo mismo. ¿Hoy también? ¿Son los macrodatos, el *big data*), tratados por la inteligencia artificial, la nueva masa de gelatina que nos convierte en un simple conjunto de números?

El teléfono móvil, dotado hoy en día con inteligencia artificial, invade nuestras vidas y también nos surgen dudas sobre lo que debe y no debe ser. Lo que sí está claro es que a finales del siglo XIX el teléfono era algo que no debía ser. Este ha seguido adelante y hemos ido superando todos los miedos y suspicacias, o quizás nosotros nos hemos ido adaptando con el tiempo.

Hoy no tenemos la sensación de perder nuestra esencia humana si hablamos por teléfono. Hemos visto que la esencia humana es más que la vida privada. Pero hoy sí pensamos que la esencia humana se ve atacada por la inteligencia artificial. Quizás quepa tener en cuenta la historia del teléfono. Con el teléfono hemos aprendido sobre nuestra naturaleza humana, con la inteligencia artificial sucederá lo mismo. En el capítulo 3 lo veremos.

De momento, damos ahora un salto gigante en el tiempo. Atravesamos corriendo el siglo XX y nos plantamos en la actualidad, con esto de la inteligencia artificial. ¿Es la inteligencia artificial lo que debe ser?

## La inteligencia artificial

### *Nos roban los trabajos*

Una de las primeras voces que alarmaron sobre la inteligencia artificial fue la de Nick Bostrom en 2014 con su libro *Superinteligencia: caminos, peligros, estrategias*[24], en el cual alertaba de los posibles peligros de una superinteligencia y pedía que esta se desarrollara con cautela. Consejo bien razonable. Nick Bostrom apenas era conocido por el público general, por lo que sus bien intencionadas ideas quedaron en el ámbito científico. Sin embargo, cuando la alarma fue retransmitida por grandes personalidades del mundo científico y tecnológico, como Bill Gates[25], Elon Musk[26] o Stephen Hawking[27], entonces nos quedamos más intranquilos.

Una de las primeras preocupaciones de la inteligencia artificial es que nos pueda suplantar en nuestros trabajos. «¿Suplantar al ser humano? ¡Eso no puede ser!». Pero calma, puede tener su explicación.

Por ejemplo, si mandamos una nave terrestre no tripulada para que corretee por los páramos de Marte, necesitamos que esté gobernada de forma autónoma por una inteligencia artificial. Supongamos que la nave terrestre estuviera dirigida desde la Tierra vía radio. Desde una sala de control se vería la situación del terreno de Marte a través de cámaras de vídeo situadas en la nave marciana. En función de lo que se viera, se podría actuar dando la orden de girar, continuar o frenar. Sería similar a volar un dron.

La diferencia con volar un dron radica en el tiempo de comunicación. Debido a la distancia entra la Tierra y Marte, el tiempo medio de comunicación entre los planetas es de unos 12 minutos. Esto significa que, si a través de las cámaras de vídeo vemos que la nave se dirige al agujero de un cráter y mandamos la orden de parar, esta llegará a la nave 12 minutos después. Pero la cosa es peor. Cuando nosotros vemos por el vídeo que la nave va hacia un cráter, eso ha ocurrido hace 12 minutos. Por tanto, desde que la nave se enfila hacia el cráter, hasta que recibe nuestra orden de frenar, habrán pasado en total 24 minutos. Para entonces, es posible que el carricoche marciano ya haya frenado por sí solo y para siempre en el fondo del cráter. En esta situación es necesario suplantar al piloto humano en la Tierra por una inteligencia artificial, que conduzca el vehículo directamente desde Marte.

De acuerdo, conducir coches por Marte necesita de la inteligencia artificial; sin embargo, este hecho no es preocupante, pues actualmente no hay muchos puestos de trabajo de chófer marciano que puedan ser sustituidos por una inteligencia artificial. La preocupación surge con el resto de actividades que sí pueden ser sustituidas por un sistema inteligente, en particular aquellas que se consideran rutinarias o peligrosas.

En este punto la situación es muy similar a lo que hemos visto con la llegada

del ferrocarril, el cual sustituyó a los cocheros de larga distancia, pero favoreció a aquellos transportes que te acercaban de la estación de tren a la población más cercana. Además, creó a su vez el empleo de conductor de tren, que era de más valor que el oficio de cochero. Toda nueva tecnología destruye puestos de trabajo y crea nuevos de más valor. Ahora bien, esto tiene un límite.

Según un informe de la OECD de 2019[28] los puestos de trabajo con menor riesgo de sustitución por una inteligencia artificial son aquellos que requieren habilidades más difíciles de automatizar, tales como la resolución de problemas, la planificación, la asesoría o la capacidad de influencia. Por ello, aquellas personas que trabajen como ejecutivos de alto nivel, directores, profesores, abogados, tecnólogos, desarrolladores de negocio, científicos o ingenieros pueden estar más tranquilas: la inteligencia artificial no se va a meter con ellos, o se va a meter menos. Sin embargo, trabajos de limpiadores, cuidadores, montadores, conductores, operadores de planta, procesadores de comida, agricultores u obreros de la construcción tienen más riesgo de sustitución.

Solución: formemos a los limpiadores, cuidadores, etc., en las habilidades difíciles de automatizar. La complejidad radica en que tales habilidades (resolución de problemas, la planificación, la asesoría o la capacidad de influencia), por el mismo hecho de ser difíciles de automatizar, son también difíciles de adquirir. Resulta complicado ver un mundo en el que todos seamos ejecutivos de alto nivel, directores, abogados, científicos o ingenieros.

Una posible solución es la renta básica universal (UBI en inglés, *Universal Basic Income*). Técnicamente consiste en disponer de una subvención mensual para todos los miembros de una comunidad, con independencia de sus ingresos y de su mérito personal, y a un nivel suficientemente alto para permitir una vida libre de inseguridad económica. Dicho en lenguaje común, significa cobrar por existir para poder vivir.

El remedio no es tan claro. Existen visiones encontradas sobre las ventajas e inconvenientes de una renta básica universal[29]. Desde un punto de vista económico se discute sobre cómo financiar tal ingreso. Es decir, ¿esto quién lo paga? También si éste favorecerá la inflación, haciendo subir los precios y, por tanto, la renta básica será insuficiente para cubrir un nivel de vida mínimo.

Se unen, además, cuestiones de ámbito moral. Cuando dependes de una subvención, pierdes parte de tu libertad, pues recibes una renta, pero dejas de tener acceso a los medios que producen tal renta. Pasas a depender de aquel que te da la pensión. También se ponen en peligro los principios de responsabilidad y reciprocidad. Cuando recibes algo a cambio de nada, no hay reciprocidad en el intercambio y se desbalancea la responsabilidad, porque pasas a tener derechos sin obligaciones. El derecho de recibir un ingreso, sin la obligación de hacer algo.

Actualmente la cuestión está en el debate académico. Con el tiempo, pasará

al debate político, pues el tema es un buen caladero de votos, y entonces primará más la ideología que el raciocinio. Hasta entonces, existen otras preocupaciones más acuciantes.

### *También se equivoca*

La inteligencia artificial, como todo sistema electrónico, es susceptible de tener brechas de seguridad y, por tanto, de ser atacado. Sin embargo, lo realmente significativo son los ataques de seguridad que buscan que el sistema inteligente se equivoque.

En un artículo de 2015 presentado por unos investigadores de Google[30] se demostraba que se podía confundir a una red neuronal en la clasificación de imágenes. Inicialmente, mostraban la foto de un oso panda, que el sistema inteligente reconocía con una fiabilidad del 57%. Posteriormente, esa foto era tratada digitalmente de tal manera que se modificaba ligeramente en su composición de píxeles. A simple vista, para nosotros, la foto seguía siendo totalmente un oso panda, sin embargo, el sistema inteligente lo reconocía como un gibón, con una probabilidad de un 90%.

El tema no pasaría de ser algo anecdótico y simpático si no fuera porque esta vulnerabilidad se podría utilizar para el crimen. Se ha demostrado que, llevando unas gafas de una montura especial con llamativos colores, se puede falsear el reconocimiento facial. Con tales gafas, una persona puede pasar por una actriz de cine a ojos de un sistema inteligente[31].

Estos temas de la seguridad, en toda innovación, son naturales y pasajeros. El tren en sus comienzos tuvo accidentes. Lo mismo ha ocurrido con la aviación y ahora ocurre con los vehículos autónomos. Con el tiempo se va mejorando la seguridad y, hoy en día, el ferrocarril y la aviación son medios de transporte suficientemente seguros, sabiendo que no existe el riesgo cero.

Así ocurre con la inteligencia artificial. Los errores de interpretación de la inteligencia artificial provocados intencionadamente se solucionan con lo que se llaman ejemplos adversos. Consiste en entrenar a los sistemas inteligentes con aquellos casos que pueden dar a error, e indicarle cuál es la solución correcta. Por ejemplo, en el caso del oso panda, el entrenamiento con ejemplo adverso consiste en enseñarle al sistema inteligente la foto ligeramente modificada del oso panda —la que confundía con un gibón— y decirle que no se confunda, que es un oso panda.

La cuestión de la seguridad es algo que nunca tiene fin. Siempre habrá personas que dedicarán toda su capacidad y buen hacer en ver cómo alterar un sistema inteligente. Frente a tales ataques, surgirán contramedidas. Entonces, esas mismas personas dedicarán ahora todos sus esfuerzos a evitar tales contramedidas. Saldrán nuevas contramedidas y así estaremos, en una historia interminable, que dará bien de comer a muchos ingenieros de seguridad.

Mientras pasamos el tiempo entre medidas y contramedidas de seguridad,

podemos ir atendiendo esas cuestiones morales que nos hablan de autonomía y verdad.

### *Despotismo digital*

Padecemos exceso de buenismo. Lo hemos visto con ferrocarril y el telégrafo. De primeras se nos muestra un mundo feliz gracias al nuevo progreso. Con el ferrocarril, los zapateros venderían más zapatos que pares de pies hay en el mundo; con el telégrafo se acabaría la barbarie; la electricidad aumentaría la felicidad de las personas, porque dispondrían de más calidad de vida. Después, la realidad es distinta.

En el caso de la inteligencia artificial, esta se pone al servicio de las redes sociales con el sano y loable objetivo de conocer al usuario para mejorar su experiencia de usuario. ¡Oh, palabra sacrosanta hoy en día! Suéltala en cualquier reunión y toda cabeza se doblará. Todo sea por la experiencia del usuario.

Mediante la mejora de la experiencia de usuario se pretende que te sientas más cómodo, más atendido y más feliz. «¡Qué buenos son conmigo! Lo hacen por mi bien». Justo ese es uno de los males más perniciosos: cuando te obligan a realizar acciones que supuestamente son para tu beneficio. Eso es el despotismo digital.

En la decadencia de las monarquías absolutistas, a finales del siglo XVIII, apareció el despotismo ilustrado. La mentalidad de aquellos monarcas despistados era la de obrar supuestamente por el bien del pueblo, pero sin contar con el pueblo: «Todo para el pueblo, pero sin el pueblo». Ahora tenemos una monarquía tecnológica, donde la inteligencia artificial es el gran valido, que nos lleva a una especie de despotismo digital: «Todo para el usuario, pero sin el usuario».

El tema puede parecer menor, pero es muy relevante. El filósofo John Stuart Mill lo trató con detalle en su obra *Sobre la libertad*[32], a mediados del siglo XIX, donde advierte sobre la tiranía de la opinión de la mayoría. Mill admite que a una persona se le pueda impedir realizar ciertas acciones, si ello supone evitar un daño a un tercero. Pero nunca está justificado obligar a alguien a realizar, o no realizar, determinados actos porque eso sea bueno para él, porque esto le haría más feliz, o porque, en opinión de una mayoría, hacerlo sería más acertado o más justo.

Mejorar tu experiencia de usuario es buscar ese supuesto bien para ti, pero sin contar contigo. De eso ya se encarga la inteligencia artificial. De pensar por ti y saber lo que de verdad te gusta. Pero no te molestes ni te enfades, que es por tu bien. ¿Cómo consigue la inteligencia artificial pensar por nosotros? Muy fácil.

Lo primero es conocer tu personalidad y para conseguirlo no hace falta una gran cantidad de información. Se ha demostrado que se puede predecir, con

una precisión entre el 60% y el 80%, tu personalidad como usuario de redes sociales simplemente analizando con *machine learning* dos elementos: tus contactos y tus publicaciones[33].

Posteriormente, se pone en juego el denominado Modelo de los cinco cactores o de los cinco grandes[34], que te clasifica de manera rápida. Según este modelo, la personalidad se compone de los siguientes cinco factores: apertura a la experiencia, en qué medida eres curioso o cauteloso; meticulosidad, si eres organizado o descuidado; extraversión, cómo eres de sociable o reservado; simpatía o tu capacidad de ser amigable, compasivo o bien insensible; y neurosis, o en qué medida tienes inestabilidad emocional.

Dependiendo de los valores en estas variables, así te comportas[35]. Si eres una persona extrovertida o abierta a las experiencias, tienes más probabilidad de usar las redes sociales, publicar más fotos o hacer más actualizaciones de tu perfil. Si tienes bajos niveles de neuroticismo —poca inestabilidad emocional— o altos valores en meticulosidad, utilizas menos las redes sociales, actualizas menos su estado y tienes menos probabilidad de tener adicción. Si eres un usuario con altos niveles en simpatía, sueles compartir fotos sobre tus actividades y visitar tu página y la de tus amigos con frecuencia.

Con este conocimiento de tu personalidad, la inteligencia artificial hace el resto: te propone recomendaciones ajustadas a eso que llaman tu perfil, que son los valores matemáticos que conforman tu personalidad. Así, si YouTube detecta que te gusta ver patochadas, conociendo tu personalidad te propone más vídeos de necedades; Netflix te sugiere el mismo tipo de película que siempre ves; o una publicación te muestra una publicidad personalizada que te invita a comprar ese producto que tú mismo ignoras que necesitas. El objetivo último es evitar que abandones las redes sociales y facturar lo máximo posible.

Este despotismo digital puede afectar a nuestros niveles de autonomía, es decir, a nuestra capacidad de tomar decisiones. John Stuart Mill, al hablar de la libertad, decía que a una persona no se le podía obligar a realizar, o no realizar, determinados actos por un supuesto bien para él. Ciertamente, Mill hablaba de obligar y a la hora de aceptar una recomendación en una red social no existe una obligación *de facto*. Lo que existe, en su lugar, es una incitación, pero una incitación tan bien creada que puede llegar a anular tu capacidad de toma de decisiones. Una seducción que se potencia por la inteligencia artificial. No te obliga, pero influye en tu capacidad de decidir. Y no nos engañemos, no es por nuestro bien, es para obtener más beneficio.

Si en su momento acabamos con el despotismo de las monarquías absolutistas, debemos evitar caer en el despotismo digital avalado por la inteligencia artificial. Está en juego nuestra autonomía.

### *Esto es de justicia*

Existe un error muy extendido en nuestra forma de hablar cuando al referirnos

a sistemas digitales decimos expresiones del estilo: «el sistema dice que usted…», «la aplicación le ha denegado su petición de…», «el sistema ha aprobado...». En estos casos aceptamos por bueno el resultado matemático de un algoritmo: «Si lo dice el sistema, entonces es que es así». Este hecho es todavía más relevante si interviene la inteligencia artificial, porque su resultado es más convincente. Pero este resultado puede no ser tan bueno, en el sentido de no ser justo o ecuánime. Esta falta de justicia puede tener dos causas: que sea por sesgos no previstos, o bien por decidir el presente en función del pasado. En los dos casos se resiente la justicia, y en el segundo, además, la posibilidad de cambiar tu vida.

Nuestra aspiración es ser justos, en el sentido de que existan igualdad de oportunidades para todos, con independencia de la cultura, sexo o ideología de cada uno. Intentamos tomar decisiones sin sesgo, ecuánimes. Como ahora las decisiones las tomamos a través de la inteligencia artificial, entonces debemos evitar que esta manifieste algún tipo de sesgo en su funcionamiento.

Uno de los primeros y más famosos casos de sesgo sucedió con Tay, un *chatbot* de Microsoft, diseñado para interactuar con jóvenes entre 18 y 24 años mediante conversaciones casuales y entretenidas a través de Twitter. Su método de aprendizaje consistía en dicha interacción con los jóvenes. A través de la propia conversación con los jóvenes se esperaba que el *chatbot* fuera aprendiendo. El resultado fue que en menos de 24 horas Tay se volvió nazi y acabó lanzando comentarios racistas y antisemitas[36]. Microsoft cerró el *chatbot* inmediatamente y los jóvenes entrenadores de Tay tuvieron algo de lo qué vanagloriarse ante sus amigos.

El caso puso de manifiesto lo importante que es entrenar a las máquinas según unos datos adecuados. De igual forma que no basaríamos la educación de un niño en la lectura indiscriminada de tuits, Twitter no es el medio más adecuado para enseñar a un sistema inteligente.

Si los datos de entrada para el aprendizaje de un sistema inteligente tienen sesgos, los resultados de ese sistema inteligente estarán sesgados. Para que el sistema ofrezca otro resultado, debe tener otros datos de aprendizaje. La inteligencia artificial no es tan inteligente como para decir, «esto que hago está mal, no estoy siendo justo, tengo sesgos» y cambiar su comportamiento. Tay no sabía que el antisemitismo era impropio de la justicia. Tan solo tenía una mayor cantidad de tuits con comentarios antisemitas y la inteligencia artificial asigna el hecho de verdad o de bueno a lo que es más abundante.

Desde entonces, cada vez más la inteligencia artificial se viene utilizando de manera rutinaria en múltiples aspectos que afectan a nuestras vidas: desde acceder a una entrevista de trabajo, obtener la libertad condicional o recibir un préstamo. Todas estas de decisiones se pueden ver afectadas por algún tipo de sesgo, lo cual puede afectar a tomar una decisión justa.

Si el sistema que recomienda candidatos para una entrevista de trabajo tiene

datos de entrada con más hombres que mujeres, recomendará candidatos varones con más probabilidad. Si al sistema que recomienda sentencias le entrenamos con datos de delincuentes reincidentes, la mayoría de ellos negros, recomendará con mayor probabilidad denegar la libertad condicional a una persona negra. Por fortuna, esta debilidad del sesgo en la inteligencia artificial está detectada y su resolución depende de dos variables: tiempo y voluntad.

Parte de los actuales sesgos provienen del propio sesgo que tienen las bases de datos de entrada de los sistemas inteligentes debido a nuestra propia historia. Por ejemplo, si en un sistema de reconocimiento de imágenes incluimos mayoritariamente fotos de cocinas donde se encuentran mujeres, porque son las fotos más abundantes que tenemos por cuestiones históricas, el sistema inteligente determinará que una persona en una foto de una cocina es una mujer[37]. Con el tiempo tendremos bases de datos más diversificadas. Fotos de cocinas con hombres y mujeres. Ya solo quedará el segundo paso, tener la voluntad de usar dichas de bases de datos sin sesgo.

La segunda causa de injusticia es quizás más problemática, pues afecta a la esencia misma de la inteligencia artificial, la cual trabaja siempre con datos del pasado, y en función de ellos intenta predecir el futuro. Pero esos datos estáticos del ayer no conciben la posibilidad de que algo pueda cambiar. Tu presente queda determinado por tu pasado. Una inteligencia artificial mal utilizada nos lleva a ser cautivos por siempre de nuestra historia. Te niega la posibilidad de cambio en tu vida. Esto se puso de manifiesto en Reino Unido en 2020 a raíz de la pandemia de la COVID-19.

Una de las muchas consecuencias de este bicho coronado fue la suspensión de las clases presenciales en los colegios e institutos durante la primera ola. La ausencia prolongada durante meses de clases, hizo que no se pudieran celebrar los exámenes de final de curso. Como solución inteligente se recurrió a la inteligencia artificial. Se decidió que la nota final de cada alumno para ese año académico se calculara por un algoritmo que tuviera en consideración dos variables: las calificaciones previas de años anteriores de dicho alumno y el rendimiento de su escuela o instituto en cuanto a la media de los resultados históricos de sus alumnos[38].

Esta decisión supuso un gran número de protestas por parte de los alumnos. No era para menos. La nota de un alumno del curso escolar 2019/2020 dependía de cuánto había estudiado ese alumno y el resto de alumnos del instituto en años anteriores. Entonces, ¿de qué le había servido estudiar ese curso? Ya se hubiera esforzado o no, su nota dependió de lo que hizo en los años previos. «Elegí un mal año para empezar a estudiar». Seguro que esto es lo que pensó un alumno que, en septiembre de 2019, antes de la pandemia, decidió cambiar de actitud y empezar a estudiar. La inteligencia artificial le negó la posibilidad de cambio. Su futuro estaba irremediablemente condicionado por su pasado.

En nuestra vida, día tras día, tenemos la posibilidad de dejar de hacer algo que quizás hemos venido haciendo hasta ese momento. Eso es ejercer nuestra libertad individual. No siempre lo practicamos, pues ejercer la libertad cuesta esfuerzo y resulta más cómodo hacer lo que siempre hemos hecho. Depende de nosotros. Pero la inteligencia artificial puede anular esa posibilidad de cambio, coartar la esperanza de que hoy no seas como eras ayer. La inteligencia artificial trabaja con matemáticas, opera con cálculos estadísticos. En las matemáticas, dos más dos siempre son cuatro; sin embargo, en nuestra vida la libertad nos ofrece la opción de sumar otro resultado.

Pero, ¡ojo!, como hemos visto en el telégrafo, depende de nosotros. La inteligencia artificial es solo una herramienta y la podemos usar de una forma u otra. En realidad, la inteligencia artificial no asignó las notas a los alumnos. Lo hicieron los responsables académicos, que en ese año de pandemia decidieron utilizar cierto algoritmo, con unos ciertos cálculos y estimaciones, para tomar decisiones sobre las vidas de los alumnos.

«Si lo dice el sistema, entonces es que es así», no es cierto. Detrás puede haber sesgos o malos usos de la inteligencia artificial. Delegar la toma de decisiones en la inteligencia artificial es acabar con la justicia. Esta nunca es el resultado de un cálculo matemático.

## Autonomía y justicia

¿Debemos tener miedo de la inteligencia artificial?

Gracias al ferrocarril íbamos a disponer de más tiempo libre, seríamos más éticos y más fraternales. Con el telégrafo acabaríamos con la barbarie. Si la velocidad de aquellos trenes decimonónicos trastornara la mente, hoy no podríamos viajar en coche. Menos en avión. Si el telégrafo impedía un pensamiento sosegado, ¿quién podría pensar hoy en día? Con el teléfono habríamos perdido toda nuestra esencia humana y hoy seríamos un montón de gelatina.

Son las predicciones, alarmantes o bondadosas, de las innovaciones del pasado. La mayor parte de ambas visiones se refutan pasados muchos años, si bien para entonces no podemos pedir explicaciones a sus autores. ¿A quién le decimos que, a pesar del telégrafo, seguimos teniendo guerras? ¿Cómo le exigimos cuentas? Esperanzar o atemorizar sale gratis, bien lo saben lo embaucadores.

Con el tiempo encontramos un nivel medio entre los deseados beneficios y las temidas amenazas. Toda innovación tecnológica nos trae un cierto beneficio y un cierto perjuicio. Pero, al final, ni los bienes son tan bondadosos ni los males tan perniciosos. Con la inteligencia artificial sucederá lo mismo.

Toda tecnología no es ni buena ni mala en sí misma, pero tampoco es neutral, porque transmite valores. Lo hemos vivido, claramente, con el telégrafo: por el valor de inmediatez que tiene, tuvimos la esperanza de unir a toda la humanidad en una comunicación instantánea, y el temor de anular por completo nuestra capacidad de pensar. Lo que debemos aprender es que el resultado del telégrafo no depende del telégrafo. Depende de nosotros. Podemos usar la comunicación para unirnos en un proyecto común o para iniciar una guerra. Igualmente, hoy en día, podemos atender notificaciones por las redes sociales o apagarlas y pararnos a pensar.

Nosotros vamos a la guerra, nosotros nos paramos a pensar. Nosotros elegimos un valor determinado de una tecnología. No es tanto buscar una inteligencia artificial ética, sino cómo podemos ser éticos y responsable con la inteligencia artificial. Más que tener miedo de la inteligencia artificial, debemos tener miedo de nosotros con la inteligencia artificial: con ella podemos ser extremadamente humanos o extrañamente fieros. Ese es su peligro, esa su fortaleza. Depende de nosotros.

Nosotros elegimos el valor de cada tecnología, dentro de los múltiples valores que esta presenta. Ahora bien, para la inteligencia artificial, ¿existe algún valor innato? ¿Existe algún verdadero riesgo ético de la inteligencia artificial?

En este repaso histórico de otras tecnologías hemos visto que muchos problemas iniciales se han ido resolviendo con el incremento de la madurez de

la tecnología. Lo mismo ocurrirá con la inteligencia artificial respecto a temas de trabajo, errores propios de la inteligencia artificial o su seguridad. Iremos resolviendo los problemas según vayamos aprendiendo y mejorando la tecnología. Pero existen una serie de riesgos que son más profundos, porque están relacionados con la propia esencia de la inteligencia artificial y los valores que transmite. Me refiero a dos valores: el valor de la autonomía y el valor de la justicia. Son los que debemos vigilar.

La inteligencia artificial posee la idea de autonomía, es decir, de poder decidir. Esto es peligroso, porque puede afectar a nuestra capacidad de decidir. Lo hemos visto con el concepto de despotismo digital, donde la inteligencia artificial piensa por ti y te ofrece lo que aparentemente es bueno para ti, pero sin contar contigo.

Luego está la idea de certeza que induce la inteligencia artificial: lo que dice la inteligencia artificial es bueno, es lo que es. Muy peligroso. Lo repito por si no he sido claro: muy peligroso. Es llevar al máximo nivel la frase «lo dice el sistema», que significa, «el resultado es verdadero». De esta forma, si la inteligencia artificial dice que esta es tu nota académica, en función de tus notas pasadas, quiere decir que es verdad, que no puedes cambiar tu vida. O si ofrece un veredicto sobre ti respecto a lo que vas a hacer, en función de lo que hiciste, este es cierto, aunque tú tengas la intención de cambiar. La idea de verdad en la inteligencia artificial afecta a nuestra libertad y a nuestra justicia.

En la historia de Hilma af Klint dije que hubo un juicio de valor erróneo y un acierto. Respecto al juicio de valor, hemos visto que nuestras visiones sobre un hecho cambian. Algo que considerábamos impropio, con el tiempo se vuelve aceptable. Si hoy vemos a la inteligencia artificial como indigna, nuestra visión cambiará. El acierto estuvo en que Kandinsky, antes de exponer una obra abstracta, explicó en su libro *De lo espiritual en el arte* cómo entender una obra de arte abstracta. Eso es lo que debemos hacer con la inteligencia artificial: entenderla, tener unos conocimientos básicos de cómo funciona. Si entendemos su funcionamiento, perderemos los miedos infundados y conoceremos sus verdaderos riesgos. Vamos a ello en el siguiente capítulo.

**Mis conclusiones. ¿Las tuyas?**

En este capítulo hemos abordado dos cuestiones: ¿Debemos tener miedo de la IA? ¿Cuáles son los verdaderos riesgos éticos de la IA? Estas son mis propuestas de respuesta:

- A lo largo de la historia, toda innovación ha tenido ilusiones y temores, centrado en aspectos de salud, seguridad, economía y moral. En esta tabla tienes un resumen de lo que hemos visto.
- Con el tiempo se ha visto que muchos temores e ilusiones eran infundados. Con la inteligencia artificial ocurrirá lo mismo.
- Toda tecnología no es ni buena ni mala en sí misma, pero tampoco es neutral, porque transmite valores. El resultado ético de la inteligencia artificial dependerá de nosotros, del valor que elijamos.
- Por ello, no es tanto buscar una inteligencia artificial ética, sino cómo nosotros podemos ser éticos y responsable con la inteligencia artificial.
- En la inteligencia artificial existen riesgos que desaparecerán, pero debemos vigilar aquellos riesgos que forman parte de los valores que transmite la inteligencia artificial: el valor de la autonomía y el valor de la justicia.
- Para evitar miedos infundados de la inteligencia artificial debemos conocer cómo funciona.
- Estas son mis conclusiones, pero no las des por ciertas. Reflexiona con autonomía y justicia para extraer tus propias conclusiones.

| | **Ferrocarril** | **Telégrafo** | **Teléfono** | **Inteligencia Artificial** |
|---|---|---|---|---|
| SALUD | La velocidad causa locura. | Nos hará perder la capacidad de pensar. | | |
| SEGURIDAD | Se producen accidentes. | | Perderemos la privacidad y la individualidad (masa informe de gelatina) | Errores propios de la IA |
| ECONOMÍA | Destruirá trabajo (transporte por canales), pero mejorará otros (zapateros). | | | Destruirá puestos de trabajo. |
| MORAL | El tiempo libre por tardar menos en viajar lo dedicaremos a ser | Acabará con la guerra. | Hablar por teléfono es indigno del ser humano. | Pérdida de autonomía.<br><br>Pérdida de justicia |

| | | |
|---|---|---|
| | más inteligentes. | (sesgos, sin libertad de cambiar). |

**Montañas y mar, Helen Frankenthaler, 1952. Fundación Helen Frankenthaler**

# Inteligencia probable

Este cuadro parece una acuarela, pero no lo es. En el verano de 1952, su autora, Helen Frankenthaler, pasó unos días en Nueva Escocia, Canadá, pintando distintas estampas de aquel paisaje costero. A su vuelta, ya en el estudio, realizó *Montañas y mar* con una técnica totalmente novedosa, que le permitió crear grandes áreas de colores difuminados, algo nunca visto hasta entonces. Nadie sabía cómo había sido capaz de crear aquellos colores tan tenues, utilizando además pintura al óleo, en lugar de acuarela. Era cosa extraña.

El cuadro causó admiración cuando lo presentó. Nadie entendía qué nueva técnica podía haber utilizado para crear aquella sensación etérea de colores y mostrar lo vaporoso de un paisaje costero. Al principio su obra no fue aceptada por la crítica. Tal y como la misma artista comentó muchos años más tarde de la presentación de la obra, *Montañas y mar* fue vista con ira, fue destrozada por la rabia. Algunos lo vieron como un gran trapo de pintura donde limpias tus pinceles, pero no algo para enmarcar[39]. Helen Frankenthaler tuvo que explicar su nueva técnica.

Habitualmente cuando se pinta un cuadro al óleo hay que preparar el lienzo. Es necesario recubrirlo de una capa que evite que la pintura entre por los poros del lienzo y, de esta forma, los colores se mantengan vivos. Es lo que se conoce como imprimar el lienzo y, justo, lo que no hizo Frankenthaler. Ella, por el contrario, dispuso el lienzo de forma horizontal en el suelo y pintó directamente en él, sin prepararlo. Además, diluyó la pintura de óleo en aguarrás, de manera que la pintura se pudiera extender y desvanecer por el lienzo.

Una vez explicada la técnica, se acabó la rabia, se acabó la ira y vinieron los reconocimientos. Helen Frankenthaler se convirtió pionera del estilo artístico llamado pintura de campos de color. *Montañas y mar* fue calificada nada más y nada menos como la piedra de Rosetta de los campos de color (la original piedra de Rosetta sirvió para el gran hito de descifrar los jeroglíficos egipcios, de ahí la comparación). Se comparó con la obra *Impresión, sol naciente* de

Monet, que dio nombre al impresionismo, porque su obra suponía una nueva etapa para el expresionismo abstracto. ¡Lo que hace explicar las cosas con tranquilidad! ¡Lo que relaja saber cómo funciona!

Se teme lo desconocido. Lo hemos visto en el capítulo anterior con el telégrafo, que estaba basado en la electricidad, la cual se consideraba como una fuerza invisible y maravillosa. ¿Cómo era posible que los mensajes llegaran de forma instantánea? Hoy el telégrafo no nos causa sorpresa, ni miedo. También lo vimos con el teléfono. ¿Cómo se puede transportar la voz por un hilo de cobre? Para dar respuesta a tan maravilloso hecho se inventaron soluciones tan peregrinas como ventriloquía o que el hilo de cobre era hueco y la voz marchaba por su interior, como por una tubería. Este desconocimiento sobre el funcionamiento del teléfono hizo que se viera como una amenaza para la persona. Hablar por un altavoz se consideró indigno del ser humano y nos avocaba a perder nuestra individualidad, a convertirnos en una masa informe de gelatina.

Lo mismo sucede con la inteligencia artificial. Nos da miedo porque no sabemos cómo funciona. Porque la llamamos inteligencia y la vemos como rival. ¡Inteligentes somos nosotros! De forma similar al telégrafo o el teléfono, surgen preguntas del estilo: ¿cómo es posible que hable?, ¿pero, cómo aprende?, ¿cómo puede escribir poemas o pintar cuadros? Y acabamos concluyendo que esto va a afectar a la dignidad humana.

En el capítulo siguiente veremos brevemente en qué sentido la inteligencia artificial se parece o no a nuestra inteligencia y si eso, por tanto, puede afectar a nuestra dignidad como seres humanos. Pero para entender eso, primero tenemos que ver cómo funciona la inteligencia artificial. El objetivo no es profundizar en detalles, no te asustes, no habrá fórmulas matemáticas, sino entender en qué principios técnicos se fundamentan los sistemas inteligentes. Para ello veremos el funcionamiento de algunas de las capacidades que podemos considerar más asombrosas de la inteligencia artificial, como puede ser el escribir o pintar, mantener una conversación, aprender o tomar decisiones.

Vamos a abrir la caja negra de la inteligencia artificial.

## Di lo más habitual

### *Hablar con un sicoanalista-máquina*

En 1950 Alan Turing publicó su famoso artículo *Computing Machinery and Intelligence*[40] donde plantea lo que denominó juego de imitación. En el juego de imitación original hay tres personas: un hombre, una mujer y un interrogador, que puede ser de cualquier sexo. El interrogador permanece en una habitación separado de los otros dos. El objetivo del juego es que el interrogador determine quién de las dos personas que están en la otra habitación es el hombre y quién la mujer. Para ello solo se puede valer de preguntas realizadas por escrito y discernir la cuestión en función de las respuestas que reciba.

Supongamos ahora que cambiamos al hombre por una máquina. La cuestión que suscita Turing es si en este caso el interrogador se equivocará o acertará a la hora de discernir quién es el hombre y quién la máquina en la misma medida que en la situación original. De esta manera, definió su famoso test de Turing, que determina que una máquina es inteligente si al hablar con ella —originalmente por escrito— creemos que estamos hablando con una persona.

Turing auguraba, en el mismo artículo, que en 50 años sería posible disponer de una máquina con la capacidad de hacer el juego de imitación de tal forma que un interrogador medio, en 5 minutos, el 30 % de las veces se equivocaría en la identificación; o bien, un 30 % de las veces la máquina pasaría el test de Turing.

Cuando Turing escribió su famoso artículo, no existía ninguna máquina capaz de hacer el juego de imitación. Lo que Turing se preguntaba es si sería posible que en algún momento hubiera alguna máquina capaz de pasar su prueba. En 1966 se creó la máquina ELIZA, que fue capaz de superarlo bajo ciertas condiciones. Podemos pensar que eso es un gran logro y que, desde entonces, las máquinas empezaron a engañarnos; pero no es así, pasar el test de Turing no es tan complicado.

ELIZA fue creada por Joseph Weizenbaum, profesor del MIT y considerado uno de los padres de la inteligencia artificial. ELIZA actuaba desde el papel de un sicoanalista que recibía preguntas y comentarios de un supuesto paciente. Este paciente es el interrogador, si pensamos en el juego de imitación, y es el que debe averiguar si está hablando con un sicoanalista o con una máquina. Toda la conversación sería por escrito. Según el propio artículo en el que se explica el funcionamiento de ELIZA, la conversación entre el paciente y el supuesto sicoanalista podría ser del siguiente estilo[41]:

PACIENTE (P): Todos los hombres son iguales.

SICOANALISTA-MAQUINA (SM): ¿En qué sentido?

P: Siempre nos molestan de una forma u otra.
SM: ¿Puedes pensar en algún ejemplo concreto?
P: Bueno, mi novio me hizo venir aquí.
SM: Tu novio te hizo venir aquí.
P: Dice que estoy deprimida la mayor parte del tiempo.
SM: Siento oír que estás deprimida.
P: Es verdad, soy infeliz.
SM: ¿Crees que viniendo aquí te ayudará a no ser infeliz?
P: Necesito ayuda, eso parece seguro.
SM: ¿Qué significaría para ti si tú tuvieras ayuda?
P: Quizás podría aprender a llevarme bien con mi madre.
SM: Háblame de tu familia.
P: Mi madre me cuida.
SM: ¿Quién más en tu familia te cuida?
P: Mi padre.
SM: Tu padre.
P: Tú eres como mi padre de alguna forma.
SM: ¿Qué parecido ves?
P: No eres muy agresivo, pero creo que no quieres que me dé cuenta.
SM: ¿Qué te hace pensar que yo no soy muy agresivo?
P: No discutes conmigo.
SM: ¿Por qué piensas que no discuto contigo?
P: Me tienes miedo.
SM: ¿Te agrada creer que te tengo miedo?
P: Mi padre tiene miedo a todo el mundo.
SM: ¿Qué más te viene a la mente cuando piensas en tu padre?
P: Amenazas.
SM: ¿Tiene eso algo que ver con el hecho de que tu novio te haga venir aquí?

La lógica que utilizó ELIZA para pasar el Test de Turing era bastante sencilla. Durante la conversación, ELIZA buscaba palabras clave en las frases del interrogador, si las encontraba, generaba una respuesta en función de reglas establecidas, por ejemplo, la palabra clave «deprimida» genera la frase «siento oír que estás deprimida»; si no las encontraba, reformulaba las palabras del interrogador, por ejemplo, «mi novio me hizo venir aquí», lo reformulaba diciendo «tu novio te hizo venir aquí»; o respondía con frases genéricas, por ejemplo, «ya veo», «por favor, continúa», que no están en el texto mostrado, pero sí en la lógica del programa.

Como se ve, no es tan complicado pasar el test de Turing. Basta con estar en un entorno muy concreto, por ejemplo, un sicoanálisis, y hablar sin decir nada. Los seres humanos somos muy buenos en esto último de hablar sin decir nada y lo hacemos en muchas ocasiones. Ello no va en contra de los objetivos del test

de Turing. Lo que Turing proponía era que, en el juego de imitación, una máquina se comportara como un ser humano y, ciertamente, los seres humanos no siempre decimos cosas inteligentes.

El objetivo último de Weizenbaum con ELIZA no era pasar esta prueba, sino investigar sobre lo fácil que es construir frases con significado semántico. En realidad, es relativamente fácil hacer que una máquina hable. Hablar es propio de los humanos y cuando vemos una máquina hablar nos quedamos sorprendidos. En 2018 Google hizo una demostración pública de cómo su asistente virtual realizaba una cita en una peluquería[42]. La persona que tomaba la cita en la peluquería no sabía que estaba hablando con el asistente de Google. En cierta forma, era un test de Turing encubierto y la demostración causó sensación. Sin ánimo de quitar mérito a los ingenieros de Google, que lo tienen, sí es verdad que no resulta tan complicado crear frases con sentido. Vamos a ver algunos ejemplos.

### *¿Quieres hablar en latín?*

Para ver lo sencillo que es crear frases nos vamos a ir a 1677. ¿1677, cuando todavía no había inteligencia artificial? Con ello veremos que los rudimentos de la inteligencia artificial son muy básicos y antiguos.

Por aquel año, el matemático inglés John Peter publicó el libro *Versificador artificial*[43] que permitía crear frases en latín con sentido. Todo lo que hacía falta era elegir un número de seis cifras, saber contar del 1 al 9 y utilizar las tablas llenas de letras que te muestro a continuación.

**Tabla 1**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| i | s | a | t | t | h | t | p | p | m | o |
| s | u | r | o | u | e | e | p | r | p | r |
| i | r | r | s | r | i | d | e | b | s | r |
| p | s | f | a | i | r | i | t | i | i | i |
| i | | d | a | d | i | d | a | m | d | b |
| a | | a | a | a | | a | a | b | | b |
| | | | b | | | | | | | |

**Tabla 2**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| d | f | f | b | i | v | v | d | d | i | a |
| a | e | u | o | e | o | a | c | c | t | t |
| r | t | r | n | m | t | t | a | l | a | a |
| b | a | n | a | a | | a | | | a | |
| a | | | b | | b | i | | b | | |

**Tabla 3**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| m | t | i | v | m | v | r | a | s | i | i |
| n | i | a | i | e | l | c | h | b | q | r |
| l | d | o | i | i | i | i | u | o | i | e |
| r | i | o | | | a | | s | s | | s |
| | b | r | m | b | | | b | | | r |
| b | | | | | | | | | | |

**Tabla 4**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| p | p | m | p | p | c | c | p | c | r | r |
| o | o | r | a | o | r | o | æ | o | n | r |
| o | u | n | o | n | d | c | s | t | m | s |
| c | d | f | i | u | t | a | i | a | e | u |
| i | c | r | r | b | t | b | d | c | r | u |
| a | a | u | t | u | u | u | m | n | n | b |
| n | u | n | n | n | a | t | t | u | t | n |
| t | t | t | n | | | n | | t | | |
| | t | b | b | t | r | | b | r | b | |
| b | b | | | | | | | | | |

**Tabla 5**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| s | s | p | f | c | t | d | i | p | i |
| o | i | œ | r | e | o | u | o | d | m |
| g | d | i | m | g | r | c | e | n | n |
| e | m | p | m | g | u | r | i | o | r |
| i | o | a | i | l | a | a | r | a | n |
| r | t | a | a | | | a | | a | a |
| a | | | b | r | | b | | | |

**Tabla 6**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| m | q | c | t | p | s | p | s | s | u | u |
| e | a | l | o | r | e | æ | l | æ | r | n |
| a | l | a | m | p | t | d | t | t | n | a |
| v | p | e | a | a | a | u | e | | a | e |
| | | m | | m | | b | | r | r | b |
| | b | | d | b | r | | | | | |

El método es asombroso. Se pueden construir casi 600.000 frases en latín, sin tener ni idea de latín. Por ejemplo, con el número 421376 se obtiene la frase «Tristia verba aliis causabunt somnia certa». A falta de conocimientos en latín, tenemos a san Google, quien con su traductor nos dice que esa frase significa «Palabras tristes encenderán ciertos sueños». O con el número 157188 se genera la frase «Pessima bella tibi producunt sidera multa», que viene a decir: «Las peores guerras han producido muchas estrellas». Si unimos los dos números ya podemos empezar a crear poesía, ¡y en latín!:

Tristia verba aliis causabunt somnia certa,
pessima bella tibi producunt sidera multa.

O bien en castellano:

Palabras tristes encenderán ciertos sueños,
las peores guerras han producido muchas estrellas.

Esto parece maravilloso, pero la ciencia que se esconde detrás de ello es bien sencilla y es la inspiración de la inteligencia artificial que hoy nos habla.

Simplemente, detrás de estas tablas con letras que parecen puestas al azar se esconden las siguientes listas de palabras:

| | **Palabras de la tabla 1** | **Palabras de la tabla 2** | **Palabras de la tabla 3** | **Palabras de la tabla 4** | **Palabras de la tabla 5** | **Palabras de la tabla 6** |
|---|---|---|---|---|---|---|
| 1ª palabra | pessima | dona | aliis | producunt | iurgia | semper |
| 2ª palabra | turpia | verba | reor | concedunt | dogmata | prava |
| 3ª palabra | horrida | vota | vides | causabunt | tempora | sola |
| 4ª palabra | tristia | iura | malis | promittunt | crimina | plane |
| 5ª palabra | turbida | bella | viro | portabunt | fœdera | tantum |
| 6ª palabra | aspera | fata | inquam | monstrabunt | pignora | certa |
| 7ª palabra | sordida | facta | tibi | procurant | somnia | quædam |
| 8ª palabra | impia | dicta | mihi | prædicunt | sidera | multa |
| 9ª palabra | perfida | damna | scio | confirmant | pocula | sæpe |

Cada cifra del número que elegimos es ahora la posición de una palabra en una lista. Por ejemplo, el número 421376 nos dice que tenemos que seleccionar la cuarta palabra de la primera tabla, después la segunda palabra de la segunda tabla, luego la primera palabra de la tercera tabla, y así hasta la última tabla. De esta forma obtenemos las palabras de la frase, que son las que aparecen en negrita.

Matemáticamente hablando, detrás del versificador artificial aparece la combinatoria. Lo que estamos haciendo es crear frases combinando palabras preestablecidas y con un orden semántico. Por ejemplo, en la lista de la tabla 2 tenemos nombres, que harán de sujeto, y en la lista de la tabla 4 tenemos verbos. Después, basta con combinar una palabra de cada lista siguiendo el orden establecido.

Estamos en 1677 y ya tenemos un sistema «inteligente» que es capaz de crear frases. Esta mecánica la podemos automatizar. Podemos pensar en una aplicación en la cual, cuando pinchas en un botón se genera un número aleatorio que posteriormente te devuelve una frase. Así tenemos un generador de frases en latín.

Pero, hoy en día ¿quién habla latín? ¿No sería más útil escribir, por ejemplo, palabras de amor?

### *Palabras de amor*

Si la naturaleza no te ha dotado con el ingenio de decir inspiradas frases amorosas, el método que se esconde detrás del versificador artificial viene en tu auxilio. Así lo pensó Christopher Strachey[44] cuando en 1954 creó su máquina de redactar cartas de amor. Si no tienes palabras para esa persona que

cuando la ves pierdes el sentido y se te cae el móvil de las manos, o llega el día de tu aniversario y no sabes qué decirle a tu amada pareja, prueba con esto:

Adorado cielo,
Tú eres mi ferviente cómplice de emociones. Mi cariño extrañamente se aferra a tu deseo apasionado. Mi empeño anhela tu corazón. Tú eres mi anhelante simpatía: mi tierno empeño.
Tuyo bellamente.
M.U.C.

M.U.C. es el nombre de la máquina, que significa Manchester University Computer, un nombre nada poético para las palabras que dice. Sea como fuere, M.U.C. hace, en esencia, lo mismo que el versificador artificial de John Peter.

El escritor de cartas amorosas, M.U.C., dispone de una serie de listas con palabras de saludos o introducción (del tipo «adorado cielo»), listas con adjetivos, nombres, verbos y adverbios, como, por ejemplo:

| Saludos 1 | Saludos 2 | Adjetivos | Nombres | Verbos | Adverbios |
|---|---|---|---|---|---|
| querido | cielo | afectuoso | devoción | anhelar | afectuosamente |
| estimado | corazón | amoroso | emoción | adorar | ardientemente |
| amado | vida | ansiado | ambición | atraer | bellamente |
| adorado | | bello | empeño | aferrar | fervientemente |
| | | anhelante | encanto | Esperar | extrañamente |
| | | ardiente | deseo | desear | cariñosamente |
| | | codiciado | afán | querer | tiernamente |
| | | ferviente | hechizo | suspirar | amorosamente |
| | | apasionado | cómplice | ansiar | tristemente |
| | | tierno | corazón | gustar | |
| | | | simpatía | penar | |

Al igual que con el versificador artificial, M.U.C. utiliza la combinatoria. La diferencia radica en que para construir las frases amorosas no basta con extraer las palabras sin más, sino que hay que introducir cierta inteligencia. Habrá que incorporar determinantes, tales como mi, mío o tú, tuyo, incluir algunos sujetos (yo, tú, nosotros), conjugar los verbos adecuadamente (en este caso, por lo general en primera o segunda persona del singular) y acomodar el género según corresponda (amado/a). M.U.C utiliza la combinatoria junto con algunas reglas gramaticales.

Algo similar, pero todavía más elaborado, hizo el ingeniero informático alemán Theo Lutz, quien en 1959 publicó en la revista *Augenblick*[45] el primer poema escrito por un ordenador. La poesía se tituló «Texto estocástico», nombre no muy poético, pero que nos da cuenta de cómo fue su origen. El poema no nació de una inspiración iluminada, sino mediante técnicas estadísticas. Los dos primeros versos del poema dicen así:

No toda mirada está cerca. No todo pueblo es tarde.
Un castillo es libre y todo granjero está lejos.
Cada extraño está lejos. Un día es tarde.
Cada casa es oscura. Un ojo es profundo.

Viendo estos cuatro primeros versos, ya podemos entender algo de la lógica que inspira el poema: parece que a la máquina le gusta la palabra tarde y que juega con el par cerca, lejos. En inteligencia artificial decir que a un sistema le gusta es equivalente a decir que tiene mayor probabilidad. Hemos incorporado la estadística en esto de crear frases.

Todo esto parece maravilloso. Nuestra capacidad de hablar y de componer frases con sentido parece que también la tienen los sistemas inteligentes y que no es algo exclusivo de los humanos como pensábamos. Quizás estemos perdiendo uno de nuestros rasgos diferenciadores. Pero la realidad es distinta. Detrás de la construcción de frases en la inteligencia artificial se encuentra la combinatoria, las reglas gramaticales y reglas estadísticas.

## Mira y aprende

### *El cerebro de Einstein*

Un sistema muy relevante dentro de la inteligencia artificial es lo que se llama red neuronal. Las redes neuronales están inspiradas en el funcionamiento de nuestro cerebro. Pretenden simular nuestra forma de pensar. Se utilizan para muchos fines, pero, por ejemplo, son la herramienta básica para reconocer objetos, es decir, mediante las redes neuronales, la inteligencia artificial es capaz de ver. Pero, ¿cómo?

Para responder a esta pregunta vamos a recurrir a Hofstadter. Me refiero a Douglas Hofstadter, científico, filósofo y académico estadounidense, y no a Leonard Hofstadter, protagonista en la serie *The Big Bang Theory* (por cierto, al igual que le pasa a Penny, la novia de Leonard en la serie, siempre me cuesta escribir este apellido de Hofstadter). En 1981 Hofstadter publicó el famoso libro *The Mind's I: Fantasies and Reflections on Self and Soul* (*El yo de la mente: fantasías y reflexiones sobre el yo y el alma*), donde aparece un curioso capítulo en el que intervienen el griego Aquiles y una tortuga[46].

La tortuga le propone a Aquiles tener todo el cerebro de Eisntein en un libro. Para ello tendríamos que recopilar datos neurona a neurona de Einstein. Cada hoja del libro sería una neurona de Einstein y en cada una de ellas guardaríamos la siguiente información:

- Un valor umbral. Que indica la predisposición a que esa página la tenga que leer. Por ejemplo, un valor umbral alto significa que va a ser difícil que lea esa página. Leer una página del libro-Einstein es similar a encender o activar una neurona de su cerebro.
- Páginas que conecta. Las siguientes páginas del libro (otras neuronas) a las cuales debo ir cuando, después de leer la página en la que estoy.
- Valores de pesos. Cantidades que me indican cómo de fácil o complicado es ir a las páginas con las cuales se conecta la neurona en la que estoy.
- Valores de cambio. Un conjunto de números que me indican cómo en un momento dado se pueden cambiar los números de las páginas con las que se conecta.

Un libro así es físicamente inmanejable. Para que nos hagamos una idea, nuestro cerebro tiene del orden de 100.000 millones de neuronas. Supongamos que el cerebro de Einstein era de tamaño similar, neurona arriba, neurona abajo. Esto significa que el libro-Einstein debería tener ese mismo número de páginas, lo que nos llevaría a un grosor de unos 20.000 km. Si dividimos el libro en dos tomos, leyendo el primer tomo llegaríamos a Japón y con el

segundo tomo, volveríamos a España. Imposible de manejar, pero un libro de estas características es el fundamento de programación de las redes neuronales.

### *Mi cerebro en números*

En nuestro cerebro una neurona se activa cuando recibe suficiente señal de todas las neuronas que le preceden y con las cuales está conectada. Es entonces cuando deja pasar su señal a otras neuronas, las cuales a su vez se activarán si reciben suficiente señal de sus conexiones. Así sucesivamente, hasta llegar a un resultado, por ejemplo, saber identificar que he visto un 7 escrito. ¿Cómo sucede esto en una red neural?

Lo primero que tenemos que tener en cuenta es que actualmente no existe una red neuronal de ámbito general que sirva para todo, como ocurre con nuestro cerebro, que sirve para hablar, para ver, para movernos, para sentir, *etc.* Depende de para qué se entrene la red neuronal. Así, por ejemplo, una red neuronal se debe entrenar para ver o para moverse. Puede ser para las dos cosas, si bien su complejidad aumenta. Es similar a tener libros-Einstein especializados: libro-Einstein para ver, libro-Einstein para moverse, *etc.*

Pero, además, tenemos la especialización dentro de la especialización. Siguiendo con la red neuronal para ver, esta identificará objetos según cómo haya sido preparada. Depende de lo que queramos que identifique. Si queremos identificar números, debemos entrenar la red neuronal para ello, y no servirá, por ejemplo, para identificar caras. Para identificar caras necesitamos preparar la red neuronal de otra forma. O bien podemos prepararla para que identifique ambas cosa, números y caras. Es similar a tener libros-Einstein de lecturas especializadas: libro-Einstein para números, libro-Einstein para ver caras, o libro-Einstein para ver ambas cosas. ¿Por qué un libro-Einstein está especializado? Por sus hojas finales. Cada libro-Einstein, cada red neuronal, es como una historia que puede tener varios finales. El libro-Einstein de identificar números, que podemos llamar libro-Einstein-ver-números, termina con diez posibles finales, con los finales del 0 al 9. Tiene diez hojas finales, con un número del 0 al 9 en cada una de ellas. Veamos entonces cómo funciona el libro-Einstein-ver-números.

Supongamos que queremos que este libro-Einstein particular vea un 7 escrito en un papel. Todo lo que tenemos para empezar es un conjunto de trazos negros sobre un fondo blanco. La disposición e intensidad de cada trazo en el papel me dice qué primeras hojas del libro debo leer. Hemos visto que en cada hoja tenemos una serie de valores: páginas que conecta, umbral (cómo de fácil se enciende la neurona), pesos (cómo de fácil se comunica con otras neuronas) y valores de cambio. Por ahora nos interesan los tres primeros valores. En cada hoja inicial de lectura, en función de una serie de cálculos con el umbral y con los pesos, obtendré a qué otras hojas del libro, con las cuales está conectada, tengo que ir. En cada una de estas nuevas hojas, unos nuevos

cálculos me llevarán a otras hojas. Así sucesivamente hasta que termine en las hojas de final de historia.

Lo curioso es que no terminaré en una sola hoja de final de historia, sino que terminaré en todas. ¡En las diez hojas finales del 0 al 9! Pero según los cálculos realizados llegaré a estas hojas finales con distinta intensidad. Depende de cómo haya dibujado el 7 en el papel: no es lo mismo dibujar un 7 claro, que un 7 que se parece a un 1 o un 9. Estas diferencias hacen que llegue a las hojas finales de historia con distinta intensidad. Tendré, por ejemplo, un final de historia 0 con intensidad 3; un final de historia 1 con intensidad 40 (el 7 dibujado se parece algo a un 1); un final de historia 2 con intensidad 30; final 7 con intensidad 80; o un final 9 con intensidad 70 (el 7 se parece algo a un 9). Todos los finales de historia, del 0 al 9, tendrán su intensidad, no los he puesto aquí por no aburrir. La intensidad significa la probabilidad. Como el final de historia 7 tiene la intensidad más alta, el libro-Einstein-ver-números concluye que lo más probable es que esté viendo un 7.

La conclusión de tanta explicación sobre el libro-Einstein es la siguiente: en una red neuronal unimos probabilidad e intención. Una red neuronal intenta simular el funcionamiento de nuestro cerebro. Si nuestro cerebro está formado por neuronas, una red neuronal está formado por unidades lógicas llamadas células, que dejan pasar una cierta señal según ciertas condiciones, hasta llegar a células finales. La red neuronal se prepara según la intención que se tenga con ella. Si nuestra intención es ver números, tendremos 10 células finales; si nuestra intención es mover un coche de juguete hacia adelante, hacia atrás, a derecha o izquierda, tendremos cuatro células finales. El resultado de la red neuronal será la célula final con más intensidad. Una red neuronal es similar a crear una novela con varios protagonistas y varios finales. Dependiendo de las primeras acciones que hagan sus protagonistas se determinará el fin más probable. Un primer paso desencadena ya un final.

La intención de una red neuronal depende de para qué se prepare dicha red neuronal. Preparar una red neuronal significa entrenarla, o, lo que es lo mismo, que la red aprenda. Entramos entonces en otro apartado misterioso de la inteligencia artificial, sobre qué es capaz de aprender. Pero primero necesitamos saber: ¿qué es aprender en inteligencia artificial?

### *Aprender es cambiar valores*

En inteligencia artificial el término aprender significa que el sistema tiene la capacidad de modificar alguna de sus valores internos. Vimos que el libro-Einstein tenía una serie de valores en cada una de sus hojas. Si estos valores permanecen sin cambio, siempre que comencemos a leer el libro por unas determinadas hojas, llegaremos a la misma historia final. Pero si conseguimos que algunos valores de cada hoja cambien, entonces al comenzar a leer el libro por las mismas determinadas hojas llegaremos a un resultado distinto. Decimos

entonces que el libro-Einstein ha aprendido. ¿Cómo conseguimos cambiar los valores escritos en cada hoja? De igual forma a como nosotros aprendemos.

Cuando éramos pequeños seguro que tuvimos libros de grandes dibujos que nos servían para identificar las cosas de nuestra vida. Seguro que veíamos un dibujo simpático de un león y nuestros padres nos decían: «esto es un león, le-ón»; o bien una oveja, y nos decían «o-ve-ja»; o siendo más mayores, veíamos un 7 y nos decían «sie-te». Lo mismo se hace con las redes neuronales.

Es lo que se conoce como fase de entrenamiento, que es el colegio de la red neuronal. Por ejemplo, a la red neuronal para identificar números se le manda al colegio de números y allí se pasa un buen rato viendo números escritos de distinta forma. Verá un montón de 7 escritos con distinto acierto, unos más claros, otros más difíciles de identificar, y en todos ellos le diremos que es un 7. Es decir, para todos ellos le decimos cuál es la historia final más probable que debe obtener. Si para un número que parece un 7 de aquella manera la red neuronal obtiene una historia final distinta, dice, por ejemplo, que es un 9, entonces cambia sus valores internos hasta llegar a la historia final de 7. Es entonces cuando ha aprendido a identificar un 7 escrito de aquella manera.

En el libro-Einstein uno de los valores que escribíamos eran los valores de cambio, que representaban cómo cambiar las hojas con las cuales se conecta cada hoja particular. Estos valores de cambio son la forma que tiene de cambiar sus enlaces a otras hojas y, por tanto, de obtener nuevos resultados, es decir, de aprender.

Este tipo de entrenamiento se llama aprendizaje supervisado, porque mandamos al sistema inteligente al colegio, donde unos profesores le enseñan. Pero en otras ocasiones la inteligencia artificial no va a la escuela, sino que aprende de la propia experiencia, acude a la universidad de la vida. Es lo que se denomina aprendizaje no supervisado. Consiste en identificar patrones de comportamiento.

Nosotros también lo hacemos. Si decimos «buenos días» y vemos que nos sonríen, entonces hemos identificado un patrón de comportamiento: decir buenos días es algo que gusta. De igual forma, una inteligencia artificial puede analizar tu comportamiento en redes sociales e identificar que cuando mandas mensajes tristes, coincide con ver muchas fotos de dulces de chocolate. Entonces el sistema ha aprendido que cuando estás deprimido te da por comer chocolate. A partir de ahora, cuando detecte algún comentario tristón, es posible que te aparezca una sugerente publicidad de crujiente chocolate. Es aprender la vida.

De nuevo volvemos a la probabilidad. Aprender en la inteligencia artificial es decirle cuál es el resultado más probable, o bien que esta lo identifique. Por eso, en el capítulo anterior hablábamos de la ausencia de justicia en el caso de las notas a los alumnos de instituto de Reino Unido. Porque la inteligencia artificial estaba dando las notas más probables para una historia previa que había

aprendido. Pero ya sabemos que las mejoras novelas son aquellas que tienen el final menos esperado, el más creativo.

Hablando de creatividad, ¿puede la inteligencia artificial ser creativa?

## Mira y copia

### *Manierismo inteligente*

En el año 2016 apareció un cuadro con el singular título de *El próximo Rembrandt*. Era el retrato de un hombre de mediana edad, con bigote y perilla, posando de lado hacia la derecha y sobre un fondo claroscuro. Vestía a la moda del siglo XIV, con capa parda, gorguera blanca en el cuello y sombrero negro de ala ancha. Todo el cuadro tiene el estilo austero de Rembrandt y su típico juego de luces y sombras. Parece un cuadro de Rembrandt, pero no es de Rembrandt. Es de la inteligencia artificial, que creó un cuadro a la manera de Rembrandt. Es una especie de manierismo inteligente.

Hacia mediados del siglo XVI surge el estilo artístico denominado manierismo, como referencia a aquellos artistas plásticos que intentaban pintar «a la manera» del gran Miguel Ángel o el singular Rafael. Su método consistía en extraer las técnicas pictóricas que estos maestros utilizaban para aplicarlas de forma mecánica. Se pensaba que la utilización reiterada de grandes técnicas llevaría a grandes obras. Ilustre error. El manierismo está considerado más como un arte decadente o de transición, que como un periodo de significativa creatividad.

La obra *El siguiente Rembrandt* fue pintada por la agencia de publicidad J. Walter Thompson Amsterdam, junto con ING, Microsoft, la Universidad Tecnológica de Delft y el Museo Mauritshuis, y siguiendo el mismo método de construcción que aquellos manieristas de finales del siglo XIV. Se diseñó a base de analizar la técnica de Rembrandt para posteriormente reproducirla. Se decidió pintar un retrato porque Rembrandt tiene una gran cantidad pinturas donde aparecen rostros de personas, lo que suponía una gran cantidad de fuentes de las que extraer su técnica.

Los diseñadores recopilaron más de 346 pinturas, que fueron analizadas por algoritmos creados para identificar patrones en su forma de pintar retratos, tales como distancia de los ojos, tamaño de la nariz, forma de la boca, vestimenta y tipos de fondo. Esto permitía después enseñar a una máquina a reproducir tales rasgos, teniendo en cuenta las proporciones finales del cuadro.

Una vez decidido el tema, se emplearon más de 500 horas de impresión para representar el retrato, que posteriormente fue completado con pintura en 3D para dar el característico volumen del óleo y las pinceladas propias de Rembrandt. Con todo ello, después de 18 meses de trastear, no tanto con pinceles y pigmentos, sino con algoritmos y fórmulas, apareció *El próximo Rembrandt*. Aplausos.

Aplausos, pero no tantos. Porque, en definitiva, este cuadro simula un Rembrandt. La inteligencia artificial es muy buena identificando patrones y esto es lo que hizo. Identificó la forma de pintar de Rembrandt, igual que

hacían los manieristas con las técnicas de Miguel Ángel o Rafael. Tecnológicamente, puede significar un hito, pero artísticamente no es muy meritorio. Toda obra a la manera de otro artista carece de interés artístico.

Escribir una novela que te recuerde a *Don Quijote de la Mancha* carece de mérito. Una vez nacido Cervantes, nadie aprecia escribir como Cervantes. Una vez nacido Beethoven, nadie aplaude una obra que suene a Beethoven. Pintar, escribir o componer como lo haría un cierto artista es fácil. Consiste en analizar las técnicas originales del autor y aplicarlas de manera mecánica, pensando que incluir mucho de algo grande es algo más grande. Por eso, utilizar la inteligencia artificial para pintar, escribir o componer como lo haría un célebre artista es relativamente fácil. Ahora bien, ¿es posible que una máquina cree una obra nueva?

Entonces Edmon de Belany levanta la mano y dice que sí.

### *Nuevo arte antagónico*

*Edmon de Belamy* es el nombre del primer cuadro totalmente original pintado por la inteligencia artificial[47]. Fue subastado en Christie's en 2018 y adquirido por un comprador anónimo por valor de unos 360.000 €. La obra es, de nuevo, un retrato, esta vez de un joven, que también posa mirando de soslayo a la derecha. Tiene un estilo más moderno, con trazos poco definidos que representan al muchacho difuminado en un fondo parcialmente oscuro. El retrato es el primero de una serie que representa a la imaginaria familia De Belamy, desde el primer conde de Belamy allá por el siglo XVIII o XIX.

La obra es novedosa y tiene un cierto estilo propio. No parece pintada a la manera de nadie. No parece manierismo digital. ¿Es posible, entonces, que la inteligencia artificial cree obras totalmente nuevas? Vamos a ver cómo fue pintado el joven Edmon de Belamy.

En el margen inferior derecho del cuadro aparece la firma de su autor, como corresponde a toda buena obra de arte. Pero en lugar de ver unos trazos que puedan simular un nombre, que habitualmente no nos dicen nada, vemos una compleja fórmula matemática que comienza por «minmax», —puede que tampoco nos diga mucho—. Es la fórmula de lo que se conoce como Redes Generativas Antagónicas (GAN en inglés, *Generative Adversarial Networks*), que son el fundamento, entre otras cosas, para crear obras de arte.

Las redes generativas antagónicas constan de dos sistemas inteligentes, habitualmente dos redes neuronales, que compiten entre sí. Para el caso de crear obras de arte, una de estas redes neuronales puede ser asimilada a un comisario de una exposición y la otra a un falso pintor. Supongamos que el comisario quiere organizar una exposición de pintura. Su objetivo es que no le cuelen ningún cuadro que no esté pintado por un merecido artista. El falso pintor, por el contrario, quiere engañar al comisario y exponer en dicha exposición. Ambos tienen objetivos antagónicos: el falso pintor quiere engañar

y el comisario quiere que no le engañen.

Este juego de engañar y no ser engañado lleva su tiempo. Para empezar, el falso pintor no tiene ni idea de pintar y comienza, por tanto, a pintar cuadros con motivos aleatorios. Cada vez que pinta un cuadro se lo presenta al comisario de la exposición. Este lo evalúa comparándolo mentalmente con obras de arte de distintos estilos y autores para determinar si el cuadro es merecedor de un gran artista. Sistemáticamente, va rechazando los cuadros por falta de calidad, pero según la cara que pone el comisario y algún que otro comentario que suelta al examinar cada obra, el falso pintor va cogiendo experiencia sobre cómo debe ser una supuesta obra maestra.

De esta forma el falso pintor va aprendiendo y llega un momento en el que tiene conocimiento suficiente como para crear un cuadro que pueda pasar por una digna obra de arte. En ese momento el pintor advenedizo ha conseguido engañar al comisario y su cuadro se colgará en las paredes de la exposición.

Desde un punto de vista técnico, el comisario de la exposición es una red neuronal llamada discriminador. Ha sido entrenada con imágenes de cuadros famosos y su objetivo es determinar la probabilidad de que una pintura sea similar a una obra de arte. El falso pintor es una red neuronal llamada generador, que está conectada al discriminador. El generador (falso pintor) le presenta obras al discriminador (comisario de la exposición) con el objetivo de conseguir que el discriminador acepte una obra suya. Comienza a generar imágenes de forma aleatoria y va perfeccionando sus obras en función de los resultados que obtiene por parte del discriminador. La red neuronal discriminador, que ha sido entrenada con una gran base de datos de obras de arte, está entrenando, a su vez, a la red neuronal generador en el arte de la pintura.

Ya vimos, al hablar del libro-Einstein, que una red neuronal puede aprender, y para ello cambia algunos de sus valores. Luego esta idea de ir perfeccionando las obras que produce el generador es posible. En particular, para pintar a Edmon de Belamy, el sistema fue entrenado con unas 15.000 imágenes de cuadros de distintas épocas[48].

El objetivo del discriminador (comisario de la exposición) es maximizar sus aciertos y minimizar los logros del generador (falso pintor). Mientras que el objetivo del generador (falso pintor) es maximizar sus logros, y minimizar los aciertos del comisario. Es lo que se llama una estrategia minimax. ¡Ahora se empieza a entender la firma del cuadro!

Esta visión de redes generativas antagónicas fue ideada por Iean Goodfellow[49] en 2014. Su apellido viene a significar algo así como buena persona o buen tipo. De ahí el apellido de la familia de Belamy, forma aproximada de «bel ami», que significa buen amigo en francés. Edmon de Belamy fue creado por la organización francesa Obivious dedicada al estudio de la inteligencia artificial y el arte.

La utilización de la técnica de redes generativas antagónicas ha evolucionado a lo que se denominan redes antagónicas creativas[50] (CAN en inglés, *Creative Adversarial Networks*). Mediante esta tecnología se han creado obras de arte de cualquier estilo: desde cuadros que representan paisajes, a cuadros abstractos con trazos o formas geométricas diversos.

Esa forma de pintar tiene dos características significativas: azar y probabilidad. La primera red neuronal, el generador, o falso pintor comienza pintando de forma aleatoria. La segunda red neuronal, el discriminador, o comisario de arte, decide si la pintura generada es una obra de arte o no. Pero en realidad, no decide. Lo que hace es asignar una probabilidad respecto a si la pintura generada es o no obra de arte. Cuando la primera red neuronal le enseña un cuadro, la segunda ofrece un resultado que dice: «esto es una obra de arte con una probabilidad de X %». Cuando esa probabilidad supera un cierto umbral definido, se considera que el cuadro es una obra de arte. Esta probabilidad es la intensidad que comentábamos en el punto anterior sobre los finales de historia de las redes neuronales.

Azar y probabilidad, una vez más y, cuando la probabilidad supera un cierto umbral, se determina una historia final y decimos que la red neuronal *decide*. Introducimos ahora el componente de tomar decisiones.

## Decide lo más deseado

### *Me pones en un brete*

Decimos que, por ejemplo, la inteligencia artificial decide cuándo mandarte una notificación al móvil, o bien decide qué ruta debe tomar para llegar a un destino, o qué hacer en un vehículo autónomo para evitar un accidente. Esto nos puede parecer maravilloso. ¿Cómo sabe la inteligencia artificial lo que debe decidir, si a veces, ni yo mismo lo sé?

Supongamos que nos encontramos en un concurso de televisión en el cual, tras numerosas y sacrificadas pruebas hemos ganado 1 millón de euros. En ese momento, antes de finalizar, viene la presentadora y nos reta: «Te propongo un juego final: te puedes llevar el millón de euros, o bien nos jugamos 3 millones de euros a cara o cruz». Vaya dilema.

Para tomar una decisión valoramos cada situación respecto a lo que me agrada o desagrada de cada una de ellas y qué nivel de certeza existe. Por un lado, tengo 1 millón seguro, con una probabilidad del 100 %. Por otro lado, tengo 3 millones con una probabilidad del 50 %, pues me lo juego a cara o cruz. Si lo que más me gusta es el dinero, o me gusta el riesgo y la aventura, me decantaré por la segunda situación, pues puedo ganar 3 millones, aunque tenga incertidumbre. Por el contrario, si me importa más la vergüenza y la cara que se me queda en el caso de que pierda (¡a ver cómo lo explico en casa!), entonces puede que me decante por la primera situación y no acepte el reto.

Estamos considerando dos factores: probabilidad y lo apetecible o deseable de cada situación. Ambos afectan. La probabilidad afecta, porque si en lugar de proponerme jugar los 3 millones a cara o cruz, me proponen que sea sacando el as de oros de una baraja española, posiblemente decida quedarme con el millón de euros que está garantizado. La probabilidad de sacar un as de oros en una baraja española es de solo el 2,5%. Pero también afecta lo deseable de cada situación. Si en lugar de prometerme 3 millones a cara o cruz, me prometen 10 millones, es posible que aceptara. En ese caso, la vergüenza de perder es menor, pues lo abultado del premio bien merece la apuesta.

En situaciones en las que interviene cierta incertidumbre tomamos las decisiones barajando la probabilidad y el deseo. Esto mismo ocurre con la inteligencia artificial. Pero la inteligencia artificial no tiene deseo En lugar de deseo se le llama utilidad, que parece un nombre más científico, y es más serio decir que la inteligencia artificial toma una decisión en función de la utilidad, que en función del deseo.

La inteligencia artificial en muchas ocasiones trabaja en lo que se llaman entornos inciertos. Son aquellos en los que la inteligencia artificial no dispone de toda la información necesaria para tomar una decisión, o bien en cualquier momento puede suceder algo que afecte al sistema. La conducción de un

vehículo autónomo sucede en un entorno incierto, pues en cualquier momento puede suceder algo en el tráfico. Un sistema de diagnóstico de una enfermedad también sucede en un entorno incierto, pero ahora porque el sistema no dispone de toda la información. Quizás el sistema inteligente está analizando una radiografía, pero no conoce todo el historial del paciente —quizás solo lo conoce el paciente y en ese momento no lo puede decir—. La inteligencia artificial que trabaja para mostrarte notificaciones en tu móvil también opera en un entorno incierto: no dispone de toda la información, pues no sabe qué estás haciendo en ese momento —afortunadamente—, y no sabe si ocurrirá algo nuevo, por ejemplo, recibir una notificación de otra red social de la competencia. En todos estos casos la inteligencia artificial toma una decisión, ya sea frenar o acelerar, diagnosticar tu enfermedad o mandarte una notificación al móvil, mediante lo que se llama la utilidad esperada. Consiste en conjugar probabilidad con deseo (utilidad). Veamos cómo lo hace.

### *La fórmula del deseo*

Cuando tenemos que decidir entre 1 millón seguro o 3 millones a cara o cruz, hacemos un cálculo informal que considera la probabilidad y lo deseable de cada situación. Este cálculo informal hay que formalizarlo de cara a la inteligencia artificial, es decir, hay que convertirlo en fórmula matemática. De esta manera surge la llamada «utilidad esperada», que es el producto de la probabilidad por la utilidad. Pero ya hemos dicho que la utilidad es en el fondo es lo deseable de una situación. Por tanto, la utilidad esperada lo podríamos llamar «deseo esperado» y es un deseo multiplicado por la probabilidad de que ocurra.

Además, la inteligencia artificial tiene una directriz muy clara al respecto: debe tomar aquella decisión que produzca una utilidad esperada más alta, el mayor deseo posible. Con esta premisa, si la inteligencia artificial estuviera en la situación del concurso de televisión, la decisión habría sido clara.

Lo primero que se le diría a la inteligencia artificial es qué debe entender por utilidad, es decir, qué es lo deseable. En la situación del concurso la respuesta es clara: lo deseable es el dinero, y cuanto más dinero, más deseable. Entonces en la primera situación tiene una utilidad esperada de 1 millón, producto de multiplicar 1 millón de euros por el 100 %. En la segunda situación tiene una utilidad esperada de 1,5 millones, producto de multiplicar de 3 millones de euros por el 50 %. Como 1,5 es mayor que 1, la inteligencia artificial decidirá jugarse los 3 millones. No hay duda.

Quizás el concursante no habría decido eso, como vimos, por el tema de la vergüenza. ¿Cómo meto la vergüenza en esta ecuación? Es más complicado. Habría que intentar cuantificar el nivel de vergüenza y meterlo en el valor de utilidad. Ahora los 3 millones de euros son menos deseables, porque hay una componente de vergüenza. Ya no multiplicaremos 3 millones de euros por

50%, sino algo menos. Si la vergüenza es muy alta, puede que la utilidad esperada de los 3 millones de euros sea menor que la del millón de euros, y opte por esto último.

La inteligencia artificial toma la decisión que le lleva a la máxima utilidad esperada, que es como decir el máximo de los deseos probables[51]. El reto radica en determinar esa idea de deseo. Si estamos en un sistema de conducción autónoma lo deseable pueden ser varias cosas. Dependerá de cada persona. Uno puede desear gastar poco combustible, otro recorrer caminos curiosos, o bien la ruta más corta; quizás alguien desee una conducción tranquila y otro conducir de forma agresiva. Pero también están los deseos de la empresa que diseña la inteligencia artificial. Quizás su deseo sea que pases por determinados sitios porque así ellos facturan por cada usuario que accede a dichos sitios. Si pensamos en un sistema de notificaciones, quizá tu deseo sea recibir notificaciones interesantes y el deseo de la red social sea mandarte muchas notificaciones para que no te vayas de su red.

A todos estos cálculos se les llama utilidad esperada. Por ello, se dirá que la inteligencia artificial tome una decisión que arroje la máxima utilidad, que es una forma de decir que la inteligencia artificial toma la decisión que produce lo que más se desea. Pero siempre nos quedará por saber cuál es esa fórmula del deseo en un doble sentido: qué es lo que se desea y quién lo desea, si los usuarios o los diseñadores de la inteligencia artificial.

Incluso ti tenemos clara la fórmula del deseo, no siempre lo que decide la inteligencia artificial coindice con nuestras decisiones. La inteligencia artificial puede *desear* una cosa y nosotros otra.

### *Decidimos por sensaciones*

La teoría de la utilidad esperada en un modelo matemático que se utiliza en el ámbito de la economía para determinar qué decisión puede tomar una persona en situaciones inciertas. Este modelo se ha traspasado a la inteligencia artificial, de tal forma que esta decide aquello con lo que obtenga una mayor utilidad esperada. Pero nosotros no somos tan racionales tomando decisiones. Esto lo demostró el economista y físico Maurice Allais al plantear la siguiente paradoja[52].

A un grupo de personas se les propuso lo siguiente, respecto a ganar un cierto dinero en una lotería. En cada opción tenemos un premio con una cierta probabilidad:

Dilema 1

Opción A: ganar 4000 € con un 80 % de probabilidad.

Opción B: ganar 3000 € con un 100 % de probabilidad.

Dilema 2

Opción C: ganar 4000 € con un 20 % de probabilidad.
Opción D: ganar 3000 € con un 25 % de probabilidad.

El primer dilema es similar al planteado en el ejemplo del concurso. Se observó que en este dilema la mayoría de las personas elegían la opción B (los 3000 euros seguros), mientras que en el dilema 2, la gente escogía la opción C. La decisión habitual sobre el dilema 1 iba en contra de la teoría de la utilidad esperada. Si en este dilema hacemos los cálculos correspondientes, matemáticamente hablando, obtenemos más utilidad esperada con la opción A, que con la opción B. Sin embargo, en el dilema 2, la decisión habitual sí coincide con lo que dice la teoría. Es decir, que en ocasiones seguimos lo que dice la teoría y en ocasiones no.

Una teoría que a veces funciona y a veces no, no es una gran teoría. Esto lo saben los economistas y los diseñadores de inteligencia artificial, por ello se realizan ajustes sobre la fórmula de la utilidad esperada. Con todo, podemos extraer una conclusión: la teoría de la utilidad esperada no dice qué debemos decidir, ni describe cómo tomamos las decisiones. Es decir, no es ni normativa ni descriptiva.

No es normativa en el sentido de que no establece cómo se deben tomar las decisiones. No hay una obligación de seguir la teoría de la utilidad esperada. Cuando nosotros decidimos, podemos seguir esa teoría o no. Como no es normativa, tampoco es descriptiva. La teoría de la utilidad no describe cómo tomamos las decisiones. Para verlo con más claridad, pensemos en otra teoría: la teoría de la gravedad. Esta sí es normativa y descriptiva, porque describe cómo cae una piedra al suelo, y la piedra siempre debe caer según dicha teoría —por ello hasta se le llama ley—.

Tomar decisiones parece algo complejo. Quizás lo sea para nosotros, porque tomamos decisiones considerando muchos factores: desde el dinero que gano hasta la vergüenza si pierdo. Para la inteligencia artificial la situación es más sencilla. Tan solo tiene que buscar la decisión que le dé el máximo deseo esperado, que en lenguaje serio se dice máxima utilidad esperada. Tanto nosotros como la inteligencia artificial tomamos decisiones en situaciones inciertas conjugando la probabilidad con el deseo, si bien de forma distinta. Nosotros lo hacemos de una manera informal, y la inteligencia artificial de forma matemática y exacta. Aquí se produce un doble alejamiento. Por un lado, nuestros deseos no siempre se ajustan a una fórmula matemática. Por otro lado, detrás de la utilidad esperada en la inteligencia artificial se encuentra una medida del deseo, pero como usuarios nunca sabemos cuál es ese deseo, ni quién desea, y por ello podemos no llegar a entender por qué la inteligencia artificial toma una decisión sobre una situación que nos afecta.

## Somos predecibles

¿Cómo funciona la inteligencia artificial?

Hemos visto cómo funcionan algunas de las capacidades de la inteligencia artificial: cómo es capaz de hablar o de componer textos; cómo las redes neuronales intentan simular nuestro cerebro y cómo deciden un resultado; cómo puede crear arte; y cómo toma decisiones en ciertas circunstancias inciertas. En todos los casos existe un denominador común: la probabilidad.

La inteligencia artificial se fundamenta en el hecho cierto de que las cosas suceden en la vida de manera más o menos uniforme o probable. Dados unos síntomas, lo más probable es una enfermedad. Esta certidumbre sobre los acontecimientos también sucede en nosotros. Somos bastante previsibles. Cuando hablamos o escribimos, lo más probable después de un sujeto es un verbo; después del verbo «ir», lo más probable es que aparezca la palabra «a» y un lugar (por ejemplo, «a casa»).

Nuestra escritura tiene un determinado estilo que hace que usemos unas palabras con más frecuencia que otras. También los artistas suelen pintar con una paleta más probable de colores, y pintan unos motivos con más probabilidad que otros. Por ello reconocemos un cuadro de Velázquez o de Goya con solo verlo. Finalmente, solemos tomar siempre las mismas decisiones. Somos muy probables tomando decisiones. El mundo y nosotros mismos somos bastante probables. Eso es lo que aprovecha la inteligencia artificial.

Una vez que la inteligencia artificial identifica patrones de comportamiento, dado que somos tan estables y probables, intervienen los diseñadores de la inteligencia artificial y le ponen un objetivo. Este objetivo es el deseo. Deseamos crear poemas de amor, recrear la charla con un sicoanalista, pintar un cuadro, identificar números o procurar que los usuarios no abandonen una red social. En esto de cumplir deseos (objetivos) la inteligencia artificial es realmente buena y seguro que lo consigue. Bastará con darle una fórmula y pedirle que calcule el valor más grande o más pequeño, que es como pedirle que nos dé el máximo de deseo, o el menor de los disgustos. Además, lo hace de forma rápida.

Ahora tenemos una explosión de la inteligencia artificial porque ahora es cuando estamos consiguiendo que sea rápida de verdad, tanto o más que nosotros. Por eso también nos asusta. Pero los fundamentos son antiguos. Lo hemos visto con el versificador artificial de John Peter de 1677, que es la base para que un sistema inteligente escriba textos.

Trabajar con probabilidades nos puede llevar a confundir lo más probable con lo cierto. Un sistema inteligente puede determinar que, si vas por una cierta ruta, te diriges entonces a un determinado destino. En realidad, lo que hará es determinar una cierta probabilidad de ir a un destino, en función de

comportamientos anteriores. Puede determinar, por ejemplo, que vas al trabajo con un 80 % de probabilidad. Esto no significa que un 80 % de tu persona vaya al trabajo. Lo cierto es que irás o no irás al trabajo. Es probable que vayas en un 80 %, pero la verdad será otra y esta será sí o no: irás o no irás; nunca irás a medias.

La probabilidad es una herramienta que nos permite trabajar con una realidad incierta, donde existen distintos grados de posibilidad. Sin embargo, la probabilidad de una realidad no es la propia realidad, en la misma medida en que la creencia no es lo mismo que la verdad. Uno puede creer que hay vida inteligente en otros planetas, y lo puede creer con una cierta probabilidad, pero luego la verdad será que haya vida o no. De igual forma, una inteligencia artificial puede creer que vas a hacer algo, estafar a un banco, o no estudiar en un curso. Asignará probabilidades en función de las veces que hayas estafado en el pasado o de cuánto hayas estudiado. Pero la verdad es que finalmente tú estafarás o no, estudiarás o no.

Helen Frankenthaler vivió la rabia y la ira contra su poética obra *Montañas y mar* porque nadie entendía su novedosa técnica de usar lienzos sin preparar y diluir la pintura en aguarrás. Lo tuvo que explicar y entonces pasó de la incomprensión al reconocimiento. Lo mismo sucedió con el telégrafo o con el teléfono. Parecían cosas misteriosas, pero ahora sabemos cómo funcionan y son tecnologías que aceptamos sin recelo. La inteligencia artificial se basa en que casi siempre hacemos lo mismo. Sorpréndela y sorpréndete haciendo algo distinto.

**Mis conclusiones. ¿Las tuyas?**

En este capítulo hemos abordado la siguiente cuestión: ¿Cómo funciona la inteligencia artificial? Estas son mis propuestas de respuesta:

- Para hacer que la inteligencia artificial hable o escriba textos se utiliza la combinatoria y la probabilidad. En el lenguaje combinamos ciertas palabras, sujetos, verbos, complementos, con una cierta probabilidad.
- Las redes neuronales ofrecen un resultado más probable entre los finales de historia que le proponemos.
- La inteligencia artificial es capaz de pintar porque identifica patrones o porque combina azar y probabilidad. Una red neuronal comienza pintando de forma aleatoria hasta llegar a una pintura que otra red neuronal determina que es una pintura con una cierta probabilidad.
- En entornos inciertos la inteligencia artificial toma una decisión mediante la teoría de la utilidad esperada, que es el cálculo del máximo deseo probable.
- Pero nosotros no somos tan matemáticos tomando decisiones, no siempre seguimos la teoría de la utilidad esperada.
- La inteligencia artificial se fundamenta en la probabilidad porque somos bastantes predecibles. Pero no siempre lo más probable es lo que luego ocurre. Depende muchas veces de nosotros.

Estas son mis conclusiones. Pero no las aceptes de forma predecible. Piensa y extrae tus propias conclusiones, y sin son distintas a las mías quiere decir que es probable otra posibilidad.

***La traición de las imágenes (Esto no es una pipa)*, René Magritte, 1928-1929.**
**Museo de Arte del Condado de Los Ángeles.**

# Esto no es una pipa

El pintor belga René Magritte sorprendió al mundo cuando pintó su famosa obra *La traición de las imágenes (Esto no es una pipa)*. El lienzo representa una pipa, perfectamente dibujada, pero debajo de la misma escribe el texto «esto no es una pipa». Desconcertante, pero cierto.

Magritte tiene razón. En verdad, lo que vemos no es una pipa. No la podemos llenar de tabaco y no la podemos fumar. Tal y como él mismo dijo «si hubiese puesto debajo de mi cuadro "Esto es una pipa", habría dicho una mentira»[53]. Lo que vemos es la representación de una pipa. Con esta obra Magritte quería destacar que lo que vemos, no es lo que es.

¿Qué ocurre con la inteligencia artificial? Si enseñáramos el cuadro de *La traición de las imágenes (Esto no es una pipa)* a un sistema inteligente de reconocimiento de imágenes, este lo clasificaría como una pipa. Podría también leer el texto que aparece debajo de la pipa y que dice «Esto no es una pipa» y almacenarlo junto con la imagen. ¿Encontraría alguna contradicción? ¿Pensaría, «esto es una pipa, pero no lo es»? Si la obra de Magritte se titula la traición de las imágenes, ¿es la inteligencia artificial la traición de la inteligencia?

René Magritte es uno de los ejemplos más representativos de lo que se denomina surrealismo. Salvador Dalí es otro ejemplo de ello, o Luis Buñuel en el caso del cine. En 1924 André Bretón escribió el primer *Manifiesto Surrealista*, donde expuso que el surrealismo buscaba expresar el funcionamiento real del pensamiento, sin la intervención de la razón[54]. Sin embargo, el término surrealismo nació unos años antes, en 1917, de la mano del poeta y dramaturgo Guillaume Apollinaire, cuando publicó la obra de teatro *Las tetas de Tiresias*, que subtituló como drama surrealista. En el prólogo de la obra, el propio autor dice que «cuando el hombre quiso imitar el caminar, creó la rueda que no parece una pierna. Así hizo el surrealismo sin saberlo»[55]. Esta relación de la rueda con el caminar es similar a la inteligencia artificial con nuestra inteligencia. Pero esto lo veremos más tarde.

El surrealismo, y la obra de Magritte en particular, trabaja sobre el

pensamiento, y eso es lo que vamos a hacer ahora. Este capítulo es algo filosófico. Su propósito es pensar sobre el pensamiento —como ves, ya empezamos con frases circulares filosóficas—. A la inteligencia artificial se le llama inteligencia y aquí te propongo entender por qué a esta tecnología se le llama inteligencia y si es igual a la nuestra.

Este capítulo tiene tres partes. En las dos primeras te invito a pensar sobre qué es pensar y cuáles son sus componentes. En la tercera parte veremos qué se entiende, en realidad, por inteligencia artificial y en qué se parece, o se diferencia, a nuestra inteligencia. Si crees que la filosofía no sirve para nada, porque no llega a conclusiones prácticas, quizás solo te interese la tercera parte. Pero si opinas, como yo, que la filosofía te ayuda a desarrollar un pensamiento crítico y a ser libre, entonces te animo a leer todo el capítulo. Recuerda, además, que uno de los objetivos de este libro es pensar. Este capítulo te invita a pensar. Pensemos, entonces si la inteligencia artificial es inteligencia, si parece una pipa, pero no es una pipa, o qué pasa con ella.

## ¿Hay alguien ahí?

### *Inteligencia significa pensar*

El término de inteligencia artificial fue acuñado por John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon en 1955. Nació con la propuesta de celebrar una conferencia en la que analizar qué es la inteligencia y ver cómo una máquina podría simularla. La conferencia se llamó Dartmouth Summer Research Project on Artificial Intelligence, más conocida como la Conferencia de Dartmouth, y se celebró el año siguiente en la universidad Dartmouth College.

El documento de propuesta de conferencia exponía los puntos que, se entendía, debía cubrir un sistema para que este se considerara inteligente. Así se hablaba de cómo programar una máquina para resolver problemas, usar el lenguaje, trabajar con conceptos y abstracciones, aprender e incorporar elementos aleatorios para generar creatividad[56]. Estas fueron las habilidades iniciales a cubrir para simular nuestro comportamiento. Otra cuestión es que se hayan cubierto o no. Lo significativo es que la inteligencia artificial nace con el propósito de diseñar máquinas que hagan lo que nosotros hacemos como seres inteligentes. Ese fue su objetivo en origen y lo sigue siendo en la actualidad. Ahora bien, ¿qué hacemos nosotros como seres inteligentes?

Aquí empieza un primer problema. Somos seres inteligentes y nos resulta complicado definir qué es la inteligencia[57]. La dificultad se encuentra en que la inteligencia no es una cosa única, sino que comprende distintas habilidades. Cuando hablamos de inteligencia nos referimos a una capacidad que nos permite razonar, planificar, resolver problemas, pensar en conceptos abstractos, comprender ideas complejas, aprender rápidamente y aprender de la experiencia[58]. Además, y esto es importante, todas estas capacidades están encaminadas a una capacidad más amplia y profunda de comprender nuestro entorno, lo que significa captar o dar sentido a lo que nos rodea.

¡La de cosas que hacemos como seres inteligentes! Algunas de ellas aparecen en la propuesta de la Conferencia de Dartmouth, tales como resolver problemas o aprender. También en la Conferencia aparece trabajar con conceptos y abstracciones. Sin embargo, en la definición de inteligencia, en lugar de trabajar con conceptos abstractos, se dice pensar en conceptos abstractos. Es un matiz relevante. ¿Qué es *pensar*?

Ahora es cuando nos ponemos filosóficos. Si queremos tener una definición de qué es pensar, podemos acudir a Descartes, famoso por su conocida frase: «Pienso, luego existo». Según esta declaración, pensar es algo muy importante, ya que está relacionado con existir. Existimos porque pensamos, y mientras tengamos pensamiento, seguiremos existiendo. Descartes decía que somos cosas que piensan, es decir cosas que dudan, conciben, afirman, niegan, quieren, no

quieren, imaginan y sienten[59]. Habla de cosas, si bien todos estos verbos exigen la presencia de una consciencia, es decir, de alguien que duda, concibe, afirma, niega, *etc.* ¡Uy!, hemos metido la palabra consciencia.

Esta idea de una consciencia lo vemos de forma más clara en otra definición de pensar, esta vez, de la mano de Kant —no podía faltar Kant—: pensar es unificar representaciones en una consciencia, o, lo que es lo mismo, pensar es emitir juicios[60]. La definición es algo más compleja que en Descartes —como no podía ser de otra forma, viniendo de Kant—, pero lo importante es entender que detrás de la idea de pensar asoma a su vez la idea de una consciencia, en el sentido de tener la capacidad de percibirse a sí mismo y en el mundo. Esto está relacionado con el final de la definición que hemos visto de inteligencia. Tenemos una serie de capacidades (razonar, planificar, resolver, aprender, etc.), pero encaminadas a comprender nuestro entorno, a dar sentido a lo que nos rodea. La conciencia entra dentro de ese dar sentido.

Pensar viene del latín *pensare*, que significa pesar. El pensamiento es lo que pesa o pondera los datos, los argumentos, las experiencias y hasta el propio pensamiento —el pensamiento puede pensar sobre lo que está pensando—. Como dice Platón en su dialogo *Teeteto*, pensar es un discurso que el alma se dirige a sí misma y el resultado de tal diálogo es un juicio sobre lo que se piensa. Aparece de nuevo la idea de pensar como una consciencia que emite un juicio.

En resumen, la inteligencia supone una serie de habilidades, tales como razonar, planificar, resolver problemas o aprender y, de manera especial, pensar. El objetivo de estas habilidades no es el conocimiento enciclopédico o ganar *Pasapalabra*, sino comprender, dar sentido a lo que nos rodea. Esta comprensión se consigue precisamente con la acción de pensar, cuyo resultado es emitir un juicio y para ello se necesita la presencia de una consciencia, es decir, de alguien que piensa, que emite un juicio. Todas estas habilidades las hacemos nosotros los humanos con nuestra inteligencia. ¿Hasta dónde llega la inteligencia artificial? ¿Piensa la inteligencia artificial?

### *Alexa te responde en chino*

En el capítulo anterior vimos que Alan Turing resolvió la cuestión de si la inteligencia artificial piensa de manera simple y apelando a la práctica. Su propuesta fue, lejos de debates filosóficos, decir que una máquina piensa si hace lo que nosotros hacemos. Punto. Esto lo podemos saber si es capaz de engañarnos y hacerse pasar por uno de nosotros. En definitiva, según Turing, una máquina piensa si pasa su llamado juego de imitación, su famoso test de Turing.

John Searle, profesor de filosofía, no está muy de acuerdo con la visión de Turing. No termina de ver el tema de la consciencia en la propuesta del test. Por ello, en 1980, treinta años después del artículo de Turing y cuando la

investigación en inteligencia artificial ya estaba dando resultados, publicó su famoso experimento mental llamado «La habitación china de Searle»[61].

Supongamos una habitación completamente aislada del exterior, salvo por una ranura a modo de boca de buzón, en cuyo interior se encuentra una máquina, que, según nos dicen, entiende chino. Para comprobarlo, por la ranura del buzón podemos introducir un texto en chino, junto con una pregunta relativa a su contenido. El texto puede ser de la forma: «Juan fue al restaurante y pidió una hamburguesa. Cuando el camarero se la trajo estaba fría. Al marcharse Juan, dejó muy poca propina». La pregunta relacionada con este texto puede ser: «¿Le gustó a Juan la hamburguesa?». Tras unos minutos de operación, por la misma ranura aparece un nuevo texto en chino que responde a la pregunta realizada. Por ejemplo, para el caso de Juan y la hamburguesa, la respuesta podría ser: «Definitivamente, no le gustó la hamburguesa y se fue disgustado del restaurante». Todo esto en chino, de tal forma que una persona versada en dicho idioma concluye, sin lugar a duda, que la máquina que está dentro de la habitación entiende chino. La máquina pasa el test de Turing en chino.

Una máquina de esta naturaleza en realidad ya fue diseñada hacia 1977[62], si bien en inglés, no en chino, siendo capaz de resolver este tipo de preguntas sencillas. Hoy en día, los actuales asistentes de voz, tipo Alexa (Amazon), Google Assistant, Siri (Apple) o Cortana (Microsoft), son una aproximación a «La habitación china de Searle». Uno puede pensar que, por ejemplo, Alexa te entiende en español o en cualquier otro idioma.

Volvamos a la habitación china. En realidad, dentro de la habitación no se encuentra ninguna máquina maravillosa. Se halla el propio Searle, el cual no tiene ni idea de chino, pero es capaz de responder a las preguntas que le hacen en dicho idioma. Para ello, Searle dispone de un completo manual que le permite operar con símbolos chinos. Por ejemplo, el manual le dice: «si encuentras el conjunto de símbolos "poca propina", entonces debes responder con estos símbolos: "no le gustó"». Entendamos que «poca propina» y «no le gustó» se encuentran escritos en chino. Como Searle no habla chino, estos textos no son más que símbolos. Searle no ve palabras con un significado para él, solo ve símbolos en chino y el manual de la habitación china le permite relacionar esos símbolos de manera adecuada. Le relaciona, por ejemplo, los símbolos «poca propina» con los símbolos «no le gustó».

La persona que se encuentra fuera de la habitación, y que hace preguntas en chino relativas a un texto en chino, está totalmente convencida de que dentro de la habitación hay alguien que entiende chino. Sin embargo, dentro de la habitación solo existe un manual y Searle, que no entiende chino. Searle plantea entonces una serie de cuestiones sobre quién entiende chino: ¿podemos decir que los manuales entienden chino?, ¿podemos decir que Searle entiende chino porque es capaz de responder preguntas en chino?, o bien, ¿es la

habitación, con el manual y Searle, quien entiende chino?

Si estas preguntas las llevamos a día de hoy, nos podemos preguntar si Alexa, Google Assistant, Siri o Cortana entienden español o cualquier otro idioma en el cual se le hagan preguntas.

En el fondo de todas estas preguntas se encuentra la idea de pensar, en el sentido de emitir un juicio, que hemos visto antes y que está relacionada con la presencia de una consciencia. Aparentemente, la habitación china parece que piensa, porque emite un juicio sobre si a Juan le gustó la hamburguesa. Parece que hay alguien que piensa en chino. Sin embargo, dentro de la habitación solo se encuentra un manual, que no piensa, y Searle, que tampoco piensa en chino. Para Searle, en este experimento, nadie piensa en chino, por tanto, pasar el test de Turing no es prueba suficiente de inteligencia. Le falta esa capacidad de pensar, de emitir un juicio con una consciencia.

Pero Searle también tiene sus detractores.

### *Una cortés costumbre*

Hofstadter no está de acuerdo con la visión de Searle. Ya sabes que me refiero al Douglas Hofstadter del que hablamos en el capítulo anterior y que publicó el libro *The Mind's I: Fantasies and Reflections on Self and Soul*, donde la tortuga le proponía a Aquiles el libro-Einstein.

Douglas Hofstadter es uno de los defensores de lo que se conoce como inteligencia artificial (IA) fuerte. Según esta teoría, una inteligencia artificial suficientemente compleja puede tener cualidades mentales similares a las humanas en todos los aspectos. Cuando la inteligencia artificial esté suficientemente desarrollada será capaz de pensar, en el sentido que hemos visto de emitir juicios, tener emociones y tener consciencia. Es una cuestión de la complejidad del algoritmo, es decir, es una cuestión de tiempo que la inteligencia artificial sea exactamente igual a nosotros. Si eso ocurre, en un momento dado tendremos replicantes que nos dirán con nostalgia que han visto arder naves de ataque más allá de Orión y que todos esos momentos se perderán en el tiempo, como lágrimas en la lluvia. Serán robots con miedo a morir.

El algoritmo en un sistema inteligente lo podemos asimilar a ese manual que se encuentra en la habitación china, que le permite trabajar con palabras en chino como si fueran símbolos. Si trasladamos la visión de Hofstadter a la habitación cde Searle, significa que, si ese manual está suficientemente elaborado, de tal forma que se pueda aplicar sobre cualquier texto y cualquier pregunta, entonces podremos decir que la habitación china pensará, tendrá emociones y consciencia como nosotros. En ese momento, de verdad entendería chino.

En particular Hofstadter está pensando en un libro similar al libro-Einstein que vimos en el capítulo anterior. Era un libro que contenía, neurona a

neurona, toda la información del cerebro de Einstein. Para Hofstadter este libro-Einstein, al ser tan complejo y contener todo el cerebro de Einstein, ya no es un libro, sino que es el propio Einstein. En este libro quedan representados sus pensamientos, sus emociones, e incluso su propia consciencia mediante los valores numéricos de cada hoja. De hecho, no tendríamos porqué llamarlo libro-Einstein, tendríamos que llamarlo directamente Einstein. El libro es Einstein y podemos referirnos a él como si fuera una persona. Tanto es así, que en realidad *leer* el libro-Einstein es hablar con Einstein.

En un momento dado de la conversación entre Aquiles y la tortuga se produce el siguiente lance respecto a cómo se debe tratar al libro-Einstein:

—Aquiles: Ahora, espera un minuto. Estás empleando el pronombre «él» sobre un proceso combinado con un enorme libro. Eso no es "él", es otra cosa […].

—Tortuga: Bueno, te dirigirías a él como Einstein […], ¿no? ¿O dirías: «Hola, libro de los mecanismos cerebrales de Einstein, me llamo Aquiles»? Creo que pillarías a Einstein desprevenido si hicieras eso. Ciertamente, se quedaría perplejo.

—Aquiles: No hay ningún «él». Me gustaría que dejaras de usar ese pronombre.

Aquiles y la tortuga ponen de manifiesto el debate, no resuelto todavía, sobre si ese libro tan complejo tiene consciencia y es el propio Einstein o no. La tortuga dice que sí, y por lo tanto no hay razón para llamarlo libro-Einstein. Si lo llamamos libro-Einstein, entonces al saludar a un amigo llamado Alfredo, tendríamos que dirigirnos a él algo así como: «Hola, conjunto de procesos cerebrales Alfredo». Por el contrario, Aquiles dice que el libro-Einstein es una cosa y no podemos tratarle como si fuera una persona.

Este libro-Einstein, como vimos, es un ejemplo de red neuronal y hoy en día no tenemos redes neuronales tan complejas. La primera red neuronal, llamada SNARC, fue creada por los estudiantes de Harvard Marvin Minsky y Dean Edmonds en 1950, y estaba formada por 40 neuronas. Actualmente, las redes neuronales pueden contener entre miles y pocos millones de neuronas. Estamos lejos, muy lejos de los 100.000 millones de neuronas de nuestro cerebro, y, por tanto, si los defensores de la IA fuerte tienen razón, estamos lejos de una posible IA fuerte capaz de tener consciencia.

Hasta entonces, hasta que llegue ese día de una inteligencia artificial suficientemente compleja, nos podemos pasar el tiempo debatiendo sobre la consciencia en los sistemas inteligentes. Unos dirán que sí, otros dirán que no, y otros pueden decir lo que argumentó Alan Turing.

En el famoso artículo «Computing Machinery and Intelligence»[63] el propio Alan Turing abordó una serie de objeciones contra su propuesta de test de inteligencia. Una de tales objeciones es relativa a la existencia de una consciencia. Su punto de vista es realmente curioso.

Alan Turing plantea que en realidad nosotros tampoco estamos seguros de que todo el mundo tenga consciencia. Si hablo con mi amigo Alfredo, damos por hecho que él es una persona como yo. Dado que yo tengo consciencia, supongo que Alfredo también disfruta de ella. Pero en verdad, a ciencia cierta no lo puedo asegurar. No puedo entrar en la cabeza de Alfredo y saber si la tiene. Como bien dice Turing: «En lugar de discutir continuamente sobre este punto, es habitual tener la cortés convención de que todo el mundo piensa».

Así se resuelve la cuestión: suponemos que todo el mundo piensa, que todo el mundo tiene consciencia, para no parecer descortés. Ciertamente nunca podremos estar seguros de si la persona con la que hablamos tiene una percepción de sí misma, de manera similar a como yo me percibo a mí mismo en el mundo. Dado que no puedo asegurar que toda persona tenga consciencia, ¿por qué negársela a una máquina? Si un replicante me habla de sus experiencias en Orión, y me dice que siente que todo eso va a desaparecer, ¿por qué no pensar que tiene consciencia como yo? Si no lo discuto con una persona, no tendría por qué discutirlo con una máquina.

Para intentar salir de este embrollo, podemos ver otro punto de vista sobre esta cuestión. Podemos pensar con arte.

## Una cuestión de arte

### *Pintar a lo burro*

No podremos aceptar que la máquina sea igual al cerebro hasta que esta sea capaz de escribir un soneto o componer un concierto en respuesta a los pensamientos y emociones que siente, y no por una cascada aleatoria de símbolos; es decir, no basta con que lo escriba, sino saber que lo ha escrito. Esta era la opinión de Geoffrey Jefferson[64], famoso neurólogo y pionero de la neurocirugía, hacia 1949. Introducimos entonces el arte en el debate. ¿Puede la inteligencia artificial crear arte?

En el capítulo anterior vimos que la inteligencia artificial pintó *El siguiente Rembrandt* y *Edmon de Belamy*. Sin embargo, no podemos decir que cumpliera las premisas de Jefferson. *Edmon de Belamy* fue pintado con azar y probabilidad y en ninguno de los dos cuadros podemos asegurar que la inteligencia artificial supiera que estaba pintando algo. Para profundizar sobre esta idea de saber que estás creando algo podemos recurrir al maestro Boronali.

Joachim-Raphaël Boronali posiblemente no sea un pintor muy conocido. Tan solo se le conoce una pintura, titulada *Y el sol se durmió en el Adriático*, fechada en 1910. El cuadro consiste en una serie de trazos de colores sin forma aparente. Una zona horizontal de tonos azules en la mitad inferior del lienzo bien podría simular un mar, y otra franja superior de colores anaranjados y amarillos pudiera ser producto de ese sol que parece dormirse en el mar. Unas pinceladas gruesas naranjas sin significado llaman la atención en un primer plano a la derecha. Con algo de imaginación uno puede ver una especie de embarcación en un mar al atardecer. La obra fue presentada en el Salón de los Independientes de París y fue vendida por 400 francos de la época (unos 3.500 euros actuales).

Para acercamos a la figura de Boronali nos debemos situar en la bohemia calle de Saules, del barrio parisino de Montmartre hacia ese año de 1910, donde se levanta el recogido y ameno café de Lapin Agile. En aquella época el café Lapin Agile era frecuentado por artistas. Allí uno se podía encontrar al mismísimo Picasso. Entre los artistas existía el debate de si aquellas expresiones en las que apenas se reconoce lo que se pinta se podrían considerar arte.

Había un escritor y periodista, llamado Dorgelès, que era contrario a las nuevas expresiones artísticas y estaba dispuesto a demostrar que el arte es algo más que pinceladas sin sentido en un lienzo. Con este propósito, en cierta ocasión tomó prestado a Lolo, el burro del dueño del café Lapin Agile, y en presencia de un funcionario judicial, para dar fe del hecho, dispuso un lienzo por detrás del jumento a una distancia adecuada, le ató una brocha a la cola del animal, y comenzó a darle zanahorias. La animación de Lolo ante el manjar que le ofrecían le hacía mover la cola con alegría, dejando así trazos de pintura

en el lienzo. Con esta inspiración tan culinaria se creó la curiosa y desconocida obra *Y el sol se durmió en el Adriático*.

Para hacer efectivo su engaño, Roland Dorgelès se inventó el nombre del inexistente Joachim-Raphaël Boronali. Tituló el cuadro como hemos dicho y lo presentó al público. No contento con ello, ideó, además, el llamado movimiento artístico el Excesivismo, con manifiesto incluido, como se merece todo movimiento artístico, encabezado por el ilustre e inexistente maestro Boronali.

Al poco tiempo, Dorgelès reveló su engaño. Señaló que su propósito era demostrar que una obra pintada por un burro podía entrar en una exposición de pintura de la época. Viendo el cuadro, uno puede dudar de que haya sido realizado completamente por un burro. Existe una foto de la época en la que se ve a un conjunto de enmascarados brindando detrás del burro Lolo y celebrando la proeza artística. El enmascaramiento solo responde a un espectáculo bohemio transgresor. En la foto se ve al jumento afanado en su tarea, lanzando algunas pinceladas sobre el lienzo, mientras degusta una zanahoria. Todo parece indicar que el pobre animal participó en parte de la obra, pero no es seguro que en toda ella. Quizás esos trazos sin forma definida que se sitúan en la parte frontal derecha puede que los hubiera realizado el inconsciente pollino.

¿Qué pinta, nunca mejor dicho, el burro Lolo en todo esto? ¿Tienen algo en común el retrato *Edmon de Belamy* y el lienzo *Y el sol se durmió en el Adriático*? Sí, lo tienen, y también con *El próximo Rembrandt*. En los tres casos tenemos a alguien que pinta un cuadro con algo; en ningún caso tenemos algo que sabe que pinta un cuadro.

Vimos que detrás de *Edmon de Belamy* se encuentra la organización Obivious y tras *El próximo Rembrand,* una serie de organizaciones. Detrás de *Y el sol se durmió en el Adriático* se encuentra un grupo de artistas provocadores encabezados por Roland Dorgelès.

En los tres casos estas personas, o grupos de personas, han decidido pintar un cuadro de forma novedosa: en dos de ellos con una inteligencia artificial, en otro caso con un pollino contento. Ni el burro ni la inteligencia artificial decidieron pintar. Habitualmente un pintor pinta con pinceles y ahora han surgido nuevas opciones. En *Edmon de Belamy* y en *El próximo Rembrandt* se han cambiado los pinceles por algoritmos matemáticos y buenas impresoras. En la obra *Y el sol se durmió en el Adriático* existen pinceles dirigidos por una mano que alimenta un burro. Si decimos que la inteligencia artificial ha pintado *Edmon de Belamy* y *El próximo Rembrandt*, tendremos que admitir que Lolo es el autor del cuadro *Y el sol se durmió en el Adriático*.

En las tres obras hay un hecho común y significativo: ni el burro Lolo ni la inteligencia artificial pusieron nombre a sus obras. Porque poner nombre a una obra de arte significa saber que la has pintado y saber que quieres decir algo

con ella. Por esta razón en ninguno de estos tres casos sus supuestos autores, la inteligencia artificial y Lolo, pusieron nombre a su obra: porque no sabían lo qué hacían. Según Geoffrey Jefferson, ni la inteligencia artificial ni el burro Lolo serían comparables a nuestro cerebro, porque no saben lo que han pintado.

¿Y qué dice de nuevo Turing a todo esto?

### *Alguien que cuenta algo*

Turing no escapa a la cuestión y la aborda desde un punto de vista sugerente, de nuevo en su famoso artículo de «Computing Machinery and Intelligence». No plantea tanto que una máquina pueda crear un soneto, cosa que da por hecho, sino que además pueda hablar sobre el propio soneto que ha escrito, lo cual supone un nivel más avanzado de criterio. Esto sería similar a lo que hemos dicho de poner nombre a un cuadro. Si tú hablas de un soneto que has escrito, quiere decir que sabes que lo has escrito y sabes lo que quieres decir con él. Turing sugiere que, quizás, en un momento dado, una máquina podría escribir un soneto y, posteriormente, tener la siguiente conversación con una persona:

PERSONA (P): En la primera línea de tu soneto que dice «te compararé con un día de verano», ¿no sería mejor decir «un día de primavera»?

MÁQUINA (M): No tendría métrica.

P: Y qué tal un «día de invierno». Esto sí qué rimaría métricamente.

M: Sí, pero nadie quiere que le comparen con un día de invierno.

P: ¿Podrías decir que Mr. Pickwick te recuerda a las Navidades?

M: En cierto modo, sí.

P: Sin embargo, las Navidades representan un día de invierno, y no creo que a Mr. Pickwick le molestara la comparación.

M: No creo que hables en serio. Por un día de invierno se entiende un típico día invernal, y no un día especial de Navidad.

Como se ve, la máquina estaría hablando sobre el contenido del propio soneto. Con este intenso diálogo Turing intenta rechazar el comentario de Jefferson cuando habla de una cascada aleatoria de símbolos. Efectivamente, esta conversación no tiene los acusados elementos aleatorios que hemos visto, por ejemplo, en el poema «Texto estocástico», en el escritor M.U.C. de cartas de amor, o en el rudimentario versificador artificial. No obstante, sí comparte con ellos los mismos elementos combinatorios y la propia visión de «La habitación china de Searle». Al final estamos hablando de combinar una serie de símbolos según unas reglas gramaticales.

Una conversación de este estilo entre una persona y una máquina es posible. Si queremos, podemos diseñar un sistema inteligente que mantenga un diálogo similar, de igual forma que se creó una máquina para ver si la hamburguesa le gustó a Juan. El algoritmo M.U.C. puede escribir cartas de amor o podemos

hacer que una máquina escriba el poema «Texto estocástico». También tenemos una inteligencia artificial para pintar cuadros o componer música. Sin embargo, todo esto no resuelve la cuestión principal que señala Jefferson, en el sentido de que no basta con escribir un texto, sino saber que lo has escrito.

A través del arte tenemos la facultad de decir algo de manera original. Detrás de una obra de arte siempre hay alguien que quiere contarnos algo. Y esto lo podrá hacer mediante pinceles, un burro, una máquina de escribir o con inteligencia artificial; pero ni el pincel pinta, ni la máquina escribe, y la inteligencia artificial tampoco. Detrás de todas las obras de arte que hemos visto creadas por la inteligencia artificial —las cartas de amor, el «Texto estocástico», *El Próximo Rembrandt* y *Edmon de Belamy*— se encuentra alguien, una persona, que nos quiere decir algo de una forma original.

Obviamente los defensores de la inteligencia artificial fuerte, Turing y Hofstadter negarán este extremo. Turing, con su diálogo sobre Mr. Pickwick, argumenta que una máquina podrá hablar sobre el poema que ha compuesto, entendiendo que lo ha escrito ella sola, sin que haya nadie detrás. Hofstadter argumentará que es cuestión de tiempo, hasta que alcancemos un algoritmo lo suficientemente complejo.

Sin embargo, la cuestión de que detrás de la inteligencia artificial siempre habrá alguien, puede tener un cierto fundamento matemático. Es el llamado problema de la decisión.

### *Sin respuesta para todo*

Curiosamente la idea de los ordenadores surgió para resolver un problema matemático planteado inicialmente en el siglo XVIII por Leibniz, quien pensó en una máquina que mediante un sistema lógico formal pudiera servir para razonar cualquier tipo de proposición. Su ilusión era poder apretar un botón de tal máquina y dejar que el sistema calculara hasta dar con la solución a cualquier problema. Según una famosa cita suya, si hubiera cualquier tipo de disputa entre dos personas, bastaría con usar esta máquina y «simplemente decir: calculemos, y sin más preámbulos, veamos quien tiene razón»[65]. Durante años se pensó si tal máquina sería posible.

David Hilbert y Wilhelm Ackermann en 1928[66] plantearon formalmente dicha pregunta con el enunciado del llamado «Problema de la decisión» (*Entscheidungsproblem*). Se preguntaban si existiría un algoritmo único que pudiera determinar si cualquier proposición era verdadera o falsa. Debería ser un algoritmo para el cual, dado un enunciado y con unas reglas de juego, fuera capaz de responder «sí» o «no» en función de si el enunciado era verdadero o falso. Es la idea de tener ese botón, en el que pensó de Leibniz, que arranca un algoritmo que me determina si lo que le pregunto es verdad o mentira. Para nuestra desgracia, o fortuna, quién sabe, no existe tal algoritmo.

Así lo demostraron, de manera independiente y mediante razonamientos

distintos, Alonzo Church[67] y Alan Turing[68] hacia 1936. La prueba de Turing tiene más relevancia para nosotros, que nos ocupamos de esto de la inteligencia artificial, dado que supuso el desarrollo de lo que conocemos como máquina de Turing y que es la inspiración de los actuales ordenadores. En cierta medida, hoy hablamos de inteligencia artificial porque hace siglos Leibniz planteó un problema matemático.

La demostración que refuta el «Problema de la decisión» es compleja. Lo que interesa es que, en definitiva, se demostró que no existe un algoritmo que pueda servir para todo. No existe la máquina de Leibniz, con un botón que resuelve cualquier tipo de problema. Para cada pregunta debo arrancar un algoritmo distinto, es decir, apretar un botón diferente. No es una cuestión de tiempo, de saber más y dar con el algoritmo adecuado que lo resuelva todo. Es, sencillamente, que no existe.

Gracias a este resultado los diseñadores de *software* tienen su futuro asegurado. Debido a que el «Problema de la decisión» es falso, somos nosotros los que decidimos qué botón apretar, es decir, qué algoritmo válido debemos diseñar para cada caso.

Roger Penrose, en su obra *La Nueva Mente del Emperador*[69], defiende que la refutación del «Problema de la decisión» nos indica que un algoritmo no puede establecer siempre la validez de otro algoritmo, no puede decidir por sí mismo. Debe existir algo externo que decida qué algoritmo utilizar para determinar la validez de algo. Ese algo somos nosotros. Nosotros somos los que decidimos a través de algoritmos.

Nosotros decidimos si pintar un retrato a la manera de Rembrandt, si crear cartas de amor, si otorgar una hipoteca a un cliente o si incitar a alguien a permanecer conectado a las redes sociales. Además, decidimos cómo hacerlo. Podremos utilizar la inteligencia artificial, pero somos nosotros los que pensamos, juzgamos y decidimos qué hacer y qué algoritmo arrancar.

## Una pipa sin consciencia de ser pipa

### *Que sea como nosotros*

Hasta ahora hemos visto lo que no es la inteligencia artificial. Corresponde ya, y más teniendo en cuenta de qué va este libro, saber qué es la inteligencia artificial.

Lo primero a considerar es que no hablamos de un único tipo de inteligencia artificial. En realidad, no tenemos una inteligencia artificial, sino que podemos decir que tenemos varias inteligencias artificiales. Vimos que el concepto de inteligencia artificial surgió de la llamada Conferencia de Dartmouth con el propósito de analizar cómo es nuestra inteligencia y cómo una máquina podría simularla. También hemos visto que nuestra inteligencia comprende varias habilidades, tales como razonar, planificar, resolver problemas, pensar en conceptos abstractos, comprender ideas complejas o aprender. Los distintos tipos de inteligencia artificial se centran en habilidades particulares de eso que llamamos inteligencia. Esto lleva a que cada inteligencia artificial tenga objetivos distintos y se fundamente en áreas de investigación específicas.

La propuesta de Turing, respecto a decir que un sistema es inteligente si pasa el juego de simulación, ha derivado en un tipo de investigación que buscar superar ese objetivo. El propósito es crear sistemas inteligentes que, bajo ciertas circunstancias, actúen de igual forma a como lo haríamos nosotros, de tal suerte que lleguemos a pensar que son personas de carne y hueso.

Es el caso del sistema ELIZA que hemos visto. El objetivo de ELIZA, que simula ser un doctor o doctora en psicología, no es ofrecer diagnósticos acertados sobre nuestra situación médica, sino hacernos creer que estamos hablando con una persona que ha estudiado psicología. Igualmente, la demostración que hizo Google sobre cómo su asistente personal inteligente era capaz de realizar una cita en una peluquería, no era para ir finalmente a cortarse el pelo. No me imagino a «Ok Google» entrando a saltitos en la peluquería para atusar su boliche sin pelo. El objetivo era decir: «Mirad lo que soy capaz de hacer, mi asistente de voz es capaz de engañarte».

En general, crear un sistema inteligente para que pase el test de Turing presenta poco interés dentro de la investigación científica de la inteligencia artificial. Está bien como divertimento en ferias y espectáculos. Tal y como argumentan reconocidos autores, como Russell y Novig[70], el gran salto de la aviación sucedió cuando se pasó de simular a las aves a utilizar los túneles de viento y aplicar las leyes de la aerodinámica. Hoy los aviones nos llevan con seguridad de un punto a otro porque el objetivo de la aeronáutica no es crear máquinas que vuelen igual que las palomas y puedan engañar a otras palomas. Científicamente hablando es irrelevante un supuesto «test de la Paloma» para la

aviación. Así sucede con el test de Turing.

No obstante, sí debemos reconocer que, con el objetivo de pasar la prueba de Turing, se han desarrollado una serie de campos específicos que han resultado beneficiosos para el desarrollo general de la inteligencia artificial. Pasar el test significa tener la capacidad de entender el lenguaje natural y de adaptarse a nuevas situaciones para poder responder en distintas circunstancias. Esto significa que es necesario investigar en las siguientes áreas de la inteligencia artificial:

- Procesamiento de lenguaje natural, para que la máquina pueda entender lo que se le dice.
- Representación del conocimiento, para almacenar de manera estructurada el conocimiento.
- Razonamiento automático, para interpretar la información que recibe y extraer conclusiones.
- Aprendizaje automático (*machine learning*), para detectar y extrapolar patrones y poder adaptarse a nuevas circunstancias.

El test de Turing inicial se planteó mediante un juego de simulación en el que la interacción entre la persona y la máquina era por escrito. Posteriormente se propuso el llamado «Total Turing Test»[71] que propone que el sistema inteligente sea además capaz de moverse como nosotros. Esto implica tener dos nuevas capacidades, como son la visión y movimiento, que dan lugar a desarrollar dos nuevas áreas de investigación:

- La percepción artificial, como, por ejemplo, el reconocimiento de imágenes.
- La traslación y actuación, para que la máquina se pueda mover en su entorno y manipular objetos.

Estas dos tecnologías se suelen aplicar a la robótica, que es otra rama de la inteligencia artificial.

Cuando hablamos de inteligencia artificial solemos hablar de todos estos temas que son componentes, o áreas de investigación de la misma. Aquí vemos cómo la idea de pasar el test de Turing nos puede servir, no tanto para superarlo con éxito, sino para desarrollar una inteligencia artificial en varios ámbitos. Así, ahora somos capaces de mandar robots a Marte, no con el propósito de engañar a supuestos marcianos y hacerles pensar que hablan con humanos, sino para estudiar Marte, mejor incluso de lo que lo haríamos nosotros. Esto nos lleva al modelo de inteligencia artificial más común que se está desarrollando.

***Obtener el mejor resultado***

La investigación de la inteligencia artificial se centra en lo que se conocen como agentes inteligentes o agentes racionales. El objetivo de esta inteligencia artificial es crear sistemas que, dado un entorno, puedan actuar para obtener el mejor resultado posible. Dos elementos: un entorno y uno, o varios, objetivos relacionados con este entorno.

Su funcionamiento se basa en los llamados agentes inteligentes, o agentes racionales. Un agente no es más que algo que hace algo en un entorno dado. Por ejemplo, esa máquina que corretea por nuestra casa barriendo el suelo, es un agente inteligente. Su objetivo es, dado un entorno (el suelo de nuestra casa), obtener el mejor resultado, que es dejarlo limpio.

Desde un punto de vista de investigación, los agentes inteligentes pueden trabajar cualquiera de los campos de la inteligencia artificial que hemos comentado en el punto anterior, si bien solo utilizan aquellos que les son estrictamente necesarios.

Así, nuestro agente barredor utiliza algunas de las áreas de investigación de la inteligencia artificial para realizar su tarea. Primero emplea razonamiento automático con reglas básicas. Reglas en el sentido de, por ejemplo: si me encuentro con una pared, me paro; si detecto agua, utilizo un tipo especial de limpieza. Dado que su objetivo solo es limpiar el suelo, no se le pide cosas muy complicadas, como, por ejemplo, analizar la basura recogida. Pero bien se podría hacer, si se quisiera, y quizás ofrecer recomendaciones a su dueño, tales como «no comas tantas patatas fritas, o no las dejes por el suelo» —no sé si nos gustaría tal robot indiscreto—. En este caso, el agente tendría un entorno, el suelo de nuestra casa, pero dos objetivos: limpiarlo y regañarnos por mala alimentación.

También utiliza las tecnologías de reconocimiento visual, para identificar obstáculos o mascotas, y la tecnología de traslación y actuación para moverse y limpiar, que es lo suyo. En este caso, el agente barredor no utiliza el procesamiento de lenguaje natural, pues, por el momento, no hablamos con él. Pero sería posible, y así, le podríamos decir «no me barras los pies, que trae mala suerte», y el agente obraría sumisamente.

Técnicamente hablando, un agente inteligente es aquel que actúa de manera inteligente. Pero esto no nos dice nada, si no aclaramos qué significa actuar de manera inteligente. En el fondo esa es la cuestión que venimos hablando en este capítulo: saber qué se entiende por inteligencia en la inteligencia artificial.

Dentro de la investigación de la inteligencia artificial, esta cuestión de actuar de manera inteligente no se basa en temas filosóficos sobre la naturaleza humana y la consciencia. Se dice que un agente actúa de manera inteligente si cumple lo siguiente[72]:

1. Sus acciones se ajustan a sus objetivos y circunstancias.
2. Es flexible frente a los cambios en el entorno y en sus objetivos.
3. Aprende de la experiencia.

4. Realiza acciones adecuadas dada su limitación de percepción y su computación finita.

Según esta definición tan concreta, nuestro agente barredor es inteligente, puesto que:

1. Sus acciones se ajustan a sus objetivos y circunstancias: me deja el suelo limpio, siempre y cuando yo no sea excesivamente guarro.
2. Es flexible frente al cambios en el entorno y en sus objetivos: si me compro un gato, lo detecta y no lo barre.
3. Aprende de la experiencia: con el tiempo puede determinar cuánto tiempo le va a llevar barrer la casa y si necesita recargar la batería o no.
4. Realiza acciones adecuadas dada su limitación de percepción y su computación finita: quizás no reconozca todo a la perfección y un día barra mi anillo de matrimonio que accidentalmente se me cayó (qué estaría haciendo), y no le pidas que reconozca obstáculos a gran distancia. No obstante, barre adecuadamente, a pesar de tener ciertas limitaciones.

¿Todo esto es inteligencia? ¿Esto que hace nuestro barredor autómata está relacionado con lo que hemos visto de inteligencia? Sí.

***No una sino varias***

Al hablar de inteligencia veíamos que esta comprende distintas habilidades: razonar, planificar, resolver problemas, pensar en conceptos abstractos, comprender ideas complejas, aprender de la experiencia. Ser inteligente no significa usar todas y cada una de estas habilidades. Nosotros no siempre estamos pensando en conceptos abstractos, y es bien notorio que no siempre aprendemos de la experiencia. Igual sucede con la inteligencia artificial basada en agentes inteligentes. Un agente inteligente no tiene que utilizar todas las habilidades que comprende la idea de inteligencia.

Ahora bien, nuestro humilde barredor utiliza varias de estas habilidades. Cuando barre, está razonando, porque está aplicando cierto razonamiento lógico, es decir, ciertos silogismos del estilo: «si me encuentro un gato, entonces parar e ir hacia atrás». Esto le sirve para resolver el tipo de problemas que se va a encontrar. También planifica, pues establece una secuencia de acciones para cumplir su objetivo de tener nuestro suelo limpio. Y, finalmente, puede aprender dónde hay zonas más difíciles, según su experiencia previa. Por tanto, nuestro barredor sin pretensiones es un agente inteligente en toda regla.

Hay algo que no hace: no piensa, en el sentido de emitir un juicio. Seguro que no está aplicando esa capacidad más amplia y profunda de comprender el entorno, de captar o dar sentido a lo que nos rodea; de valorar, sopesar y

juzgar. Seguro que no está pensando «¡qué guarro es mi dueño!».

Tenemos, por tanto, varios tipos de inteligencia artificial, según nuestro objetivo de investigación. Una de ellas puede estar encaminada a simular nuestro comportamiento, y llegado el caso, pasar el test de Turing. Otra se fundamenta en agentes inteligentes, cuyo objetivo es, dado un entorno de actuación, obtener el mejor resultado posible. Existen además distintas tecnologías que ayudan a desarrollar estos tipos de inteligencia artificial: procesamiento de lenguaje natural, representación del conocimiento, razonamiento automático, aprendizaje automático (*machine learning*), percepción —por ejemplo, reconocimiento de imágenes—, traslación y actuación. Cada tipo de inteligencia artificial utilizará todas o parte de estas tecnologías, según sea su objetivo. A modo de ejemplo, en la tabla siguiente tienes una serie de sistemas de inteligencia artificial con las tecnologías que utilizan.

| | Procesamiento del lenguaje natural | Representación del conocimiento | Razonamiento automático | Aprendizaje automático | Percepción artificial | Traslación y actuación |
|---|---|---|---|---|---|---|
| **Vehículo autónomo** | X | X | X | X | X | X |
| **Pago a través del móvil por reconocimiento facial** | | | X | X | X | |
| **Sistema de recomendación de contenidos** | | X | X | X | | |
| **Sistema de diagnóstico médico por imágenes** | X | X | X | X | X | |
| **Chatbot de resolución de dudas** | X | X | X | X | | |
| **Barredor de casa** | | | X | X | X | X |

Estas tecnologías se fundamentan en una serie de disciplinas de conocimiento tales como: la filosofía, que, a través de la lógica, y, en particular, de los silogismos, ofrece principios que ayudan a la demostración de sentencias; las matemáticas y, de manera especial, la estadística, a través de las cuales podemos definir algoritmos para la resolución de problemas; la economía, que ofrece fórmulas que ayudan a determinar la toma de decisiones —como vimos en el capítulo anterior al hablar de la función de utilidad—; la neurociencia, con el estudio de las neuronas, que permite desarrollar las redes neuronales; la sicología, que permite entender cómo pensamos; la ingeniería computacional, que se ocupa de diseñar programas cada vez más eficientes, más rápidos y que

consuman menos recursos; o la llamada cibernética, que define mecanismos para que un sistema se autocontrole, por ejemplo, para que vaya a velocidad constante.

Habitualmente se toma la visión de los agentes inteligentes como la más apropiada para el progreso de la inteligencia artificial. Este enfoque es más completo, porque incluye a todos los demás, y es más práctico, porque no busca solo parecerse a los seres humanos, sino que persigue un ideal de actuación y razonamiento. No importa, por ahora, si tiene consciencia.

### *Ni falta que hace*

¿Es la inteligencia artificial igual a nosotros?

Nuestra inteligencia representa una serie de habilidades, tales como razonar, planificar, resolver problemas, aprender o pensar. De todas estas habilidades, la más especial es pensar, porque significa emitir un juicio y esto necesita la presencia de una consciencia. Por ello surge la cuestión sobre si las máquinas con inteligencia artificial piensan. Alan Turing lo resuelve de forma práctica mediante el juego de simulación o test de Turing. Según este juego, una máquina es inteligente si es capaz de hacer lo que nosotros hacemos y nos puede engañar, o imitar. Con esto, evita la cuestión filosófica sobre si la máquina piensa o emite juicios y si tiene una consciencia. Según su propuesta, basta con que parezca que emite juicios.

Magritte es puro surrealismo. Es el surrealismo que en el prólogo de la obra de teatro *Las tetas de Tiresias* de Apollinaire proclama que «cuando el hombre quiso imitar el caminar, creó la rueda que no parece una pierna».

La inteligencia artificial comparte, o debe compartir, esa misma idea surrealista. Si queremos mejorar el caminar, creamos la rueda, no nuevas piernas; si queremos mejorar el volar, creamos aviones, no nuevos pájaros. De igual forma, si queremos mejorar nuestra capacidad de obrar en el mundo, no es necesario crear una tecnología que haga lo que nosotros hacemos, sino crear una tecnología que sea muy buena en aquello en lo que nosotros no lo somos tanto, por ejemplo, en calcular rápidamente.

Cuando Magritte pintó una pipa, debajo escribió «esto no es una pipa». Tenía razón. Aquello no era una pipa, sino la representación de una pipa. De igual forma, la inteligencia artificial no es exactamente como nuestra inteligencia y no tiene por qué serlo. La inteligencia artificial no es capaz, al menos por ahora, de pensar, en el sentido de comprender una situación, valorarla y juzgarla. Ni falta que hace. Para eso estamos nosotros, quienes juzgaremos, mejor o peor, pero no creo que la inteligencia artificial sea capaz de juzgar mejor que nosotros. La justicia no es cuestión de matemáticas y cálculos estadísticos.

El objetivo no es que en un momento dado la inteligencia artificial nos llegue a engañar y se haga pasar por uno de nosotros —otra cosa es que otros

busquen engañar con la inteligencia artificial—. El objetivo de la inteligencia artificial, con esa visión de los agentes inteligentes, es obtener el mejor resultado en un entorno dado. De igual forma que una rueda o un avión me dan un mejor resultado. Y no me preocupa si la rueda o el avión se parecen a mi capacidad de andar o de moverme por el aire.

Quizás la polémica de la inteligencia artificial es que tiene la palabra *inteligencia*, y eso nos recuerda a nosotros. Es el cuadro de una pipa que no es una pipa. Si quiero fumar, no me compro un cuadro que representa una pipa. De igual forma, si quiero pensar, no debo acudir a máquina que simula pensar. Podré acudir a ella como apoyo, como complemento. Para pensar, estamos nosotros.

El arte nos distinguirá de la inteligencia artificial —con permiso de Turing y Hofstadter—. Detrás de una obra de arte siempre hay alguien que quiere contar algo. Ya sea mediante pinceles, un burro, o mediante inteligencia artificial. Detrás de la inteligencia artificial siempre hay alguien que toma la decisión de cómo hacer algo. Nosotros elegimos el algoritmo. Nosotros tenemos la posibilidad de huir de los patrones, de entender y dar sentido a nuestro entorno y ser capaces de crear algo único que diga:

Me gustas cuando callas porque estás como ausente,
y me oyes desde lejos, y mi voz no te toca.
Parece que los ojos se te hubieran volado
y parece que un beso te cerrara la boca.

*Veinte poemas de amor y una canción desesperada.*

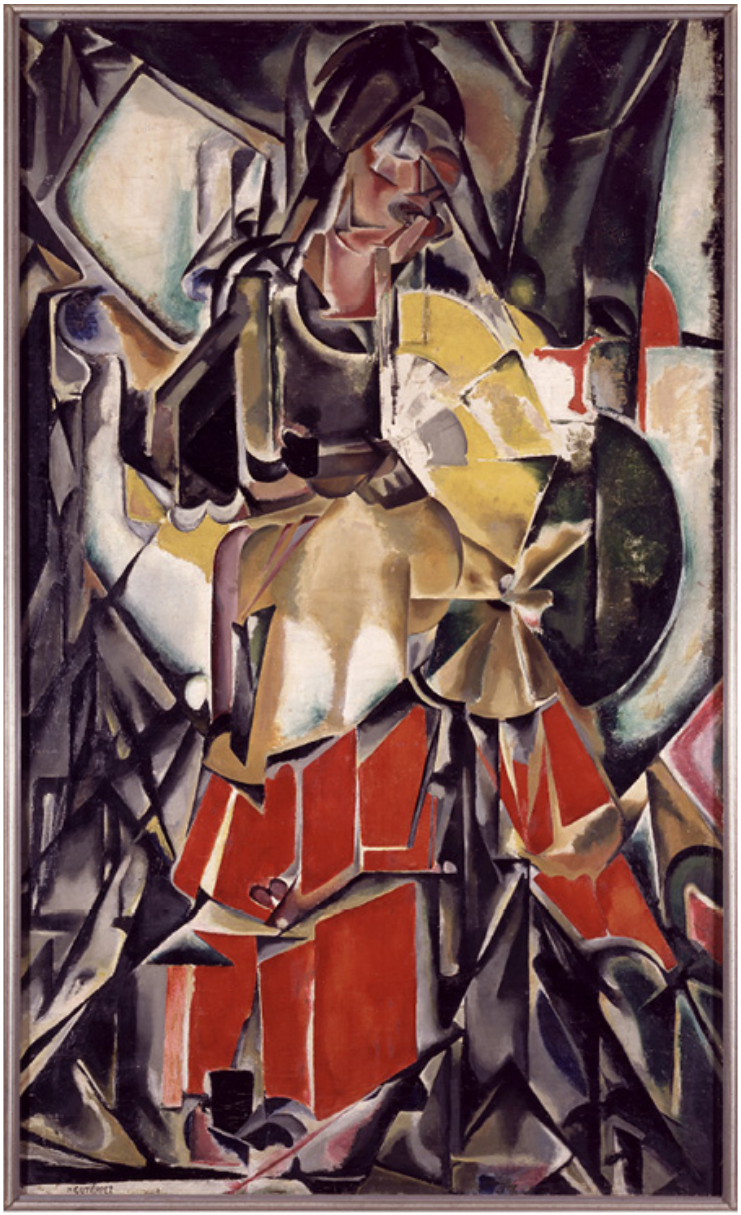
Pablo Neruda

**Mis conclusiones. ¿Las tuyas?**

En este capítulo hemos abordado la siguiente cuestión: ¿Es la inteligencia artificial igual a nosotros? Estas son mis propuestas de respuesta:

- Nuestra inteligencia es una capacidad que nos permite razonar, planificar, resolver problemas, pensar en conceptos abstractos, comprender ideas complejas, aprender rápidamente y aprender de la experiencia. Todas estas capacidades están encaminadas a una capacidad más amplia y profunda de comprender nuestro entorno, lo que significa captar o dar sentido a lo que nos rodea.
- Esta capacidad de dar sentido a lo que nos rodea está relacionada con pensar, en el sentido de emitir un juicio.
- Pensar, a su vez, necesita de una consciencia, que es la capacidad de percibirse a sí mismo y en el mundo.
- Existen distintas visiones sobre si la inteligencia artificial tiene o podrá tener consciencia. La inteligencia artificial fuerte (Hofstadter) defiende que es un tema de complejidad de los algoritmos, y en un momento dado, tendrá consciencia. Searle, con su propuesta de Habitación China defiende que la inteligencia artificial no sabe lo que hace y solo trabaja con símbolos. Turing dice que si no nos preguntamos si los demás tienen consciencia y lo damos por hecho, por qué negarlo para la inteligencia artificial.
- El Problema de la Decisión nos indica que detrás de una inteligencia artificial siempre hay alguien que toma una decisión respecto a qué hacer y cómo hacerlo.
- La inteligencia artificial actual se fundamenta en los llamados agentes inteligentes, que son aquello que actúan en un entorno para obtener el mejor resultado.
- Estos agentes son inteligentes porque sus acciones se ajustan a sus objetivos y circunstancias, son flexibles frente al cambios en el entorno y en sus objetivos, aprenden de la experiencia, realizan acciones adecuadas dada sus limitaciones. Se les considera inteligentes por hacer esto, no se busca, por ahora, que tengan consciencia.
- La inteligencia artificial utiliza distintas tecnologías.

Estas son mis conclusiones. Pero sé inteligente y piénsalas, en el sentido que hemos visto de emitir un juicio. Júzgalas y quizás obtengas otras.



***Mujer con abanico*, María Blanchard, 1916. Museo Nacional Centro de Arte Reina Sofía**

# Será por éticas

—¡Pues lo que haga más feliz al usuario! Todos buscamos la felicidad y tenemos que transmitir la idea de que esa felicidad es posible. ¡Con esa visión nació nuestra empresa!

Todo el Comité de Dirección se quedó callado tras escuchar las palabas de Ariadna, directora ejecutiva de la *startup* AlegrIA. La empresa comenzó a rodar hacía apenas dos años con el propósito de ofrecer los mejores contenidos multimedia a sus clientes mediante una potente plataforma basada en inteligencia artificial. Aquella mañana el Comité de Dirección estaba reunido para revisar el criterio del algoritmo de inteligencia artificial que recomendaba nuevos contenidos a los usuarios.

La sesión se inició compartiendo cada uno cómo se sentía emocionalmente en ese momento. Así lo indicaban las últimas buenas prácticas de gestión de reuniones, las cuales incidían en la importancia de contar a tus compañeros cómo te sentías, para crear un vínculo afectivo favorable al comienzo de cada reunión. Todo el mundo manifestó sentirse contento, ilusionado, a tope, retador o de cualquier otra forma que significaba eso que llaman energía positiva, salvo Nieves, quien manifestó sentirse algo cansada por pasar mala noche. Sin reparos, pero de forma guay, el resto del comité la animó a ver la situación desde una perspectiva positiva, cosa que Nieves no terminó de entender, quizás por el sueño que tenía, quizás por la tontería de propuesta.

La reunión se inició como una tormenta de ideas, donde cada cual podía expresar lo que quisiera, sin ser juzgado por ello. A pesar de tanta energía positiva y tan buena intención, los ánimos se fueron caldeando, al igual que en las reuniones tradicionales, llegando un momento en el que había más tormenta que ideas. Fue cuando intervino Ariadna, con las palabras que hemos visto. Se hizo un silencio momentáneo, que pronto fue roto para ser respondida.

—Vale, pero qué tipo de felicidad, porque esto depende de los gustos de cada uno —replicó Humberto con autosuficiencia.

—Buscar la felicidad de cada uno es complicado. Mejor ofrezcamos lo que le

gusta a la mayoría y así hay menos riesgo de equivocarnos. Digo yo —propuso, Benito, mientras se encogía de hombros, como queriendo no meter ruido.

—Bueno, ¡vamos a ver! Esto tampoco puede ser lo que quiera cada uno o lo que quiera la mayoría. ¡Así nos va! —saltó Tomás, con su gesto de complacencia habitual siempre que hacía una crítica a la sociedad actual—. De alguna manera todos sabemos que lo que está bien o no. Apliquemos el sentido común, ¡por favor!

—Pues partamos entonces de unos principios. ¡Tiene que haber calidad! —replicó con energía Kanza.

—Un poco de orden, ¡por favor! —gritó Habana, levantando las palmas de las manos buscando concordia—. Tendremos que buscar una aprobación de la mayoría, un diálogo entre todas las partes, con los usuarios, los productores de contenidos…

—Pero, ¿quiénes somos nosotros para determinar lo que está bien o lo que está mal? —irrumpió Nieves, quien siempre mantenía una visión escéptica de la vida.

En ese momento surgió la tormenta perfecta, todos hablando a la vez, apoyando sus ideas, sin que nadie escuchara a nadie. En un extremo de la mesa se encontraba un joven, más joven que los miembros del Comité de Dirección de AlegrIA, que tenía la dificultosa labor de tomar notas de la reunión, para luego levantar acta y compartirla con todos los presentes. Era el becario. Hasta ahora aquella labor había sido más o menos posible. Ahora resultaba imposible en aquella batalla verbal por obtener la razón a gritos.

Aprovechando que tenía bajo su control el proyector de vídeo, en un momento del acalorado debate, quitó la presentación que había servido de base para revisar la estrategia de contenidos, y proyectó una imagen de un cuadro que había estado buscando por Internet. El cuadro era *Mujer con abanico*, pintura cubista de la artista santanderina María Blanchard. Dejó de tomar nota y esperó.

Al cabo de unos segundos, se hizo el silencio cuando todos se percataron de que un cuadro raro había sustituido la notable presentación de la directora. Ariadna espetó:

—¡Esto qué es! ¿Has tomado nota de todo lo que hemos dicho?

—Esto es la solución —respondió el becario con una sonrisa abierta.

Obviamente, el becario fue despedido de AlegrIA por comportamiento inesperado. Nieves dijo que aquello del cuadro se podría considerar una visión innovadora, pero fue acallada bajo la crítica unánime de que no estaba lúcida por falta de sueño. El becario tenía razón en su propuesta.

La *startup* AlegrIA es ficticia, y los diálogos de su supuesto Comité de Dirección también. Sin embargo, todos los argumentos que hemos visto son frases que hoy en día podemos escuchar en cualquier tipo de debate o discusión. Nos pueden parecer novedosos y hacernos creer que nuestro

pensamiento es moderno. Pero todos esos argumentos han sido expuestos en los siglos y siglos de filosofía que llevamos, desde los tiempos de los griegos. Todo lo que pensamos hoy en día, ya ha sido pensado en el pasado, y, lo que es mejor, ya ha sido discutido, aprobado y rebatido, por lo que conocemos las ventajas e inconvenientes de cada idea que propongamos. Esta es la gran ventaja de conocer la filosofía, que no pierdes el tiempo, porque lo que escuchas ya lo conoces.

La historia de la filosofía, al menos en lo que respecta a la ética, es similar a un hilo de Twitter, donde continuamente se proponen y rebaten propuestas de teorías. En un momento dado, un filósofo propone una teoría, que siglos después es discutida y rebatida por una nueva teoría de otro filósofo, que siglos después es discutida y rebatida por otra teoría de otro filósofo, que siglos después… Así hasta nuestros días. La diferencia con Twitter es que, en la historia de la filosofía el debate es respetuoso y racional.

Por eso, hoy en día, tenemos tantas visiones sobre lo que está bien y lo que está mal y no llegamos un consenso sobre una visión única. Tenemos un cuadro ético cubista, como el cuadro *Mujer con abanico*, que presentó el becario de AlegrIA. La pintura representa una mujer, con falda roja, que parece llevar un abanico amarillo en su mano izquierda. La imagen de la mujer está descompuesta en distintos puntos de vista. Como rota o fragmentada en distintos planos. Así es el cubismo. Se abandona la visión de la realidad desde una perspectiva única y se intenta plasmar en un lienzo plano toda la riqueza de puntos de vista que tiene el espacio en el que nos movemos.

El cubismo nació como respuesta a la diversidad. En particular cuando artistas, como Picasso, se encontraron con el arte africano, que ofrecía imágenes de formas simplificadas y con una visión totalmente distinta de la cultura tradicional renacentista. Aquel nuevo arte era otro punto de vista. Esa misma diversidad de puntos de vista la tenemos en la ética actual, tras siglos y siglos de debate, y, por ello, la moral de nuestros días nos parece un cuadro cubista difícil de ver. Un cuadro en el que cada uno tiene su opinión sobre lo que está bien o está mal.

Esta diversidad de visiones se traslada también al ámbito de la inteligencia artificial. Si queremos crear una inteligencia artificial ética, surge la primera cuestión, que parece sin respuesta aparente: ¿qué ética aplicamos en la inteligencia artificial? Porque no tenemos un cuadro ético definido; tenemos un cuadro ético cubista formado por muchos puntos de vista, por muchas teorías éticas.

Eso es lo que le ocurría al Comité de Dirección de AlegrIA. Había demasiadas teorías éticas en juego. Los nombres de sus miembros no son casuales. Sus primeras letras hacen guiños a distintos filósofos. Ariadna es de la escuela aristotélica —no en todo—; Humberto se guía por los sentidos, como Hume; Benito es utilitarista, como Bentham; Tomás sigue a santo Tomás —

este es fácil—; Kanza es kantiana, busca principios, como Kant; Habana se inspira en el discurso práctico de Habermas; y Nieves lo rompe todo, como hizo Nietzsche con la ética.

Cada uno de ellos es *follower* de un filósofo, y no es consciente de ello. Habla creyendo que es original, sin saber que su pensamiento está subscrito a la cuenta Instagram de un filósofo. Así sucede hoy en día a todos. Existe la falsa creencia de que la filosofía son «pájaros y flores», y lo que dice no se puede aplicar a la vida real. Sin embargo, los argumentos de cada uno de los miembros de AlegrIA son «realmente reales», tan reales como que los hemos dicho o escuchado más de una vez. Y todos ellos son la consecuencia de una corriente filosófica.

¿Cómo entender y aplicar este cuadro ético cubista? Si tenemos un cubismo ético, ¿es posible tener una inteligencia artificial ética? La respuesta es sí. Vamos a verlo. Primero, vamos a profundizar en estas distintas perspectivas éticas, que hemos visto en las distintas voces del Comité de Dirección de la *startup*., porque representan teorías éticas relevantes de la historia de la filosofía, con sus pros y sus contras. Además, lo vamos a ver siguiendo el objetivo que tenía el Comité de Dirección de AlegrIA, que era revisar el algoritmo de recomendación de contenidos.

Actualmente, se utiliza la inteligencia artificial en muchos sistemas que sirven para ofrecer contenidos; ya sea, por ejemplo, de películas en Netflix, de historias en Instagram o de notificaciones en LinkedIn. Te propongo ver qué dirían estos filósofos si les preguntáramos cómo hacer que la inteligencia artificial para la propuesta de contenidos fuera ética. Con esto, entenderemos de verdad el cuadro cubista que tenemos. Para nuestra tranquilidad, en el siguiente capítulo, daremos la solución a cómo ver un cuadro cubista.

## ¿Qué es esto de la ética?

### *Unas primeras preguntas éticas*

Nosotros nos comportamos con unas reglas que no siempre son únicas. Si una leona caza y mata a su presa, la respuesta de por qué actuó así es clara y única: porque tenía hambre. Y cada vez que tenga hambre, cazará y matará. Pero nosotros podemos tener hambre y actuar de distintas formas: podemos matar o robar por conseguir comida; podemos trabajar por conseguirla; si conseguimos comida, podemos comerla toda o compartirla con otros; o podemos ayunar. ¿Por qué podemos actuar de distintas formas? ¿es alguna de ellas mejor que otra? ¿es alguna la correcta, la que «debe ser»? Estas preguntas tan habituales son preguntas éticas, principalmente la última, que se resume en: ¿cómo debo actuar? Las teorías éticas intentan responder a tales preguntas desde distintos puntos de vista.

Una primera teoría ética, que se considera como el origen en sí de la ética, se encuentra en el diálogo de Platón llamado *Protágoras*. En él se cuenta un mito con el cual se pretende dar respuesta a de las preguntas anteriores[73]:

Una vez que Zeus creó la tierra y todos los seres vivos, se dio cuenta de que tenía que dotarlos de recursos para que pudieran sobrevivir. Para ello recurrió al titán Epimeteo a quien envió a la tierra para distribuir a todas las especies distintas facultades. De esta forma, por ejemplo, a unos animales les dotó de velocidad, a otros de fuerza, a otros les dio alas para volar y a otros la habilidad para guarecerse en cualquier escondrijo. Distribuyó así todas las facultades y por eso hoy tenemos tantos animales con capacidades tan diversas. Pero Epimeteo no era muy precavido y se gastó todos los dones en los animales de tal forma que no le quedó ningún recurso que dar a la especie humana.

Prometo, su hermano, se percató de tal hecho e intentó solucionarlo. Robó el fuego a Hermes y las distintas artes de obrar a Atenea. El fuego lo dio para conocimiento de todos los humanos y las distintas artes las repartió entre distintos seres humanos de forma individual. De esta forma, los seres humanos nos especializamos en distintas tareas. Algunos recibieron el arte de hacer calzado, otros de hacer ropas, aquellos de cultivar la tierra, construir naves o edificios.

Parecía que Prometeo había resuelto el error de su hermano. El problema surgió, precisamente, cuando se crearon las ciudades y los seres humanos se juntaron en ellas. Entonces se empezaron a pelear unos con otros. Habíamos recibido muchos dones, pero nos faltaba el sentido de la convivencia.

Tuvo que intervenir Zeus para salvar a la especia humana. Envió a Hermes para distribuir entre los seres humanos dos dones adicionales: el sentido del honor y el sentido de la justicia. Antes de partir, Hermes le preguntó a Zeus cómo debía repartir tales dones. Si debía ser como las anteriores artes, que se otorgaron de forma individual, o bien debía dotarlos a todos por igual en el sentido del honor y la justicia. La respuesta de Zeus fue clara: debía repartir estos dos dones entre todos por igual, pues si participaran de ellos

solo unos pocos, como ocurría con las demás artes, jamás habría ciudades. Además, daba la facultad a los hombres para poder expulsar de la ciudad a aquellos que no participaran del honor y la justicia.

Esta historia resuelve varias cuestiones. Primero, ya sabemos por qué tenemos fuego y por qué existen distintos oficios. Después, se explica por qué los seres humanos tenemos una idea de qué es justo y por qué tendemos a esa justicia. El don de la justicia nos dice qué es lo correcto, lo que debe ser; mientras que el don del honor es ese sentimiento de vergüenza que podemos tener si no hacemos eso que es justo o correcto. Según la historia de Prometeo, sabemos qué es lo justo y nos da vergüenza ante los demás no realizar tal justicia. Ese sentimiento de vergüenza es el origen de la moral.

Por último, este mito aporta una cuestión adicional: no todos podemos opinar sobre cómo hacer zapatos, pero todos podemos opinar sobre cómo ser justos. Este punto es interesante porque pasa de soslayo en el texto, pero defiende esta idea tan actual de que en el ámbito de la ética todo es opinable.

Las artes de obrar que Prometeo robó a Atenea, las repartió de forma individual. De esta forma, unos son zapateros, otros sastres, otros agricultores y otros arquitectos. Si queremos saber cómo se hace un zapato, preguntamos a los zapateros y solo a los zapateros, pues solo ellos saben de zapatos. Solo los zapateros pueden opinar sobre cómo hacer zapatos. Sin embargo, el don de la justicia se repartió entre todos los seres humanos. Esto significa que todos los seres humanos sabemos sobre la justicia. Por tanto, todos podemos opinar sobre qué es o no es justo. Ahora entendemos por qué en los debates de televisión acuden y opinan tertulianos de toda índole para hablar de lo bueno y lo malo: resulta que aplican el mito de Prometeo, quizás sin saberlo.

Esta conclusión, algo simpática, pone de manifiesto lo práctico de la filosofía. Puede parecer que hablamos de historietas teóricas, que no resuelven nada, pero no es verdad. Al comienzo de este apartado nos hacíamos una serie de preguntas, a raíz de nuestro comportamiento respecto a de una leona: ¿por qué podemos actuar de distintas formas? ¿es alguna de ellas mejor que otra? ¿es alguna la correcta, la que «debe ser»? Este mito de Prometeo responde a tales preguntas.

¿Por qué podemos actuar de distintas formas? Los animales recibieron dones y actúan según tales dones. La leona tiene la fuerza y velocidad para cazar. Nosotros recibimos el don de la justicia que nos hace obrar según las circunstancias.

¿Es alguna de ellas mejor que otra? ¿es alguna la correcta, la que «debe ser»? Sí, y esa forma correcta la determina el propio ser humano, porque todos podemos opinar sobre qué es justo y además podemos expulsar de la ciudad a quien consideremos que no cumple tal idea de justicia.

Como vemos, detrás de cada teoría ética aparece una idea sobre la realidad y sobre el ser humano, que tiene su efecto práctico. Para el mundo griego según

este mito el ser humano puede opinar sobre cualquier tema. Esto se traduce en la famosa frase, que seguro hemos escuchado de que «el hombre es la medida de todas las cosas». Hoy en día, pensamos de igual forma, si bien desde fundamentos distintos.

Este texto de Prometeo se considera como una primera teoría ética: tenemos el don de la justicia y queremos hacer justicia porque una divinidad, Zeus, nos ha dado el don de entender qué es justo y de tener vergüenza si no lo hacemos; además, explica por qué todos podemos opinar sobre qué es justo. Esta teoría responde a la cuestión de cómo sabemos lo que es justo y por qué intentamos hacer el bien y la justica, pero, inevitablemente, levanta otras cuestiones: ¿si no creo en dioses, de dónde viene ese don?, ¿se puede discutir sobre la justicia de la misma forma que se discute sobre cómo hacer zapatos?; si hay unas normas para hacer buenos zapatos, ¿cuáles son las normas para ser justo?; si todos podemos opinar sobre esas normas, ¿para qué las normas?

La historia de la ética es la historia de un intento de responder a este tipo de cuestiones. Una teoría ética responde a unas cuestiones de manera más o menos razonada. Pero, como dije antes, al responder a algunas preguntas, nacen otras nuevas preguntas. Viene entonces una nueva teoría que responde a las nuevas preguntas, pero genera otras tantas, o bien es una teoría no satisfactoria para todos y nace una nueva teoría ética. Así, llegamos al debate del Comité de Dirección de AlegrIA y al cuadro cubista ético. Lo iremos viendo con más detalle. Ahora terminemos de cerrar qué es esto de la ética.

### *Vida buena con justicia*

Hemos visto que el mito de Prometeo es una primera teoría ética que explica por qué tendemos a un comportamiento justo. Sin embargo, si nos fijamos bien, esta historia no dice nada sobre cómo nos tenemos que comportar para ser justos. Explica por qué tenemos el don de la justicia, pero no cómo comportarse para ser justo. Esta es la diferencia entre la *ética* y la *moral*, como ahora veremos.

Habitualmente las palabras ética y moral se entienden como sinónimos y se utilizan de manera indistinta. Parte de tal confusión se debe al origen de estas palabras. *Ética* procede del griego *ethos*, que significaba originariamente «morada», «lugar donde vivimos», y que posteriormente pasó a considerarse como «el modo de ser de una persona»; algo así como la morada particular de cada uno. Por otro lado, *moral* procede del latín *mos*, *moris*, que originariamente significaba «costumbre», pero que evolucionó también a «modo de ser una persona»; sería como la costumbre particular de cada uno. De este modo, tanto *ética* como *moral* confluyen en la idea de todo aquello que se refiere al modo de ser de una persona. Modo de ser que se supone ha sido adquirido como resultado de poner en práctica una cierta costumbre o unos hábitos.

Según esto, ética y moral son lo mismo y, en cualquier conversación,

podemos utilizarlas según queramos. No obstante, por si se da el caso de que nos encontramos con alguien puntilloso y preciso, que se siente estupendo en cualquier momento, vamos a profundizar en cómo se diferencian ambos conceptos.

La diferencia radica en qué tipo de preguntas nos hacemos respecto a ese modo de ser o de obrar de una persona. Cuando yo quiero realizar una acción me pueden venir dos tipos de preguntas: una, ¿puedo hacer eso? y dos, ¿por qué debo o no debo hacer eso? La primera pregunta la responde la moral, y la segunda pregunta la responde la ética.

Volviendo al caso de tener hambre, me puedo preguntar dos cosas: ¿puedo robar comida?, o bien, ¿por qué debo o no debo robar comida? La moral responde a la cuestión de si robar o no robar, si bien la respuesta depende de cada tipo de moral. Por ejemplo, la moral cristiana dirá «no robarás», mientras que la moral llamada utilitarista —lo veremos posteriormente— dirá «depende de si robas para hacer el bien o robas para hacer el mal» —de ahí el aprecio romántico que tenemos por Robin Hood, que robaba para el bien de los pobres—. La ética, sin embargo, responde a la pregunta de por qué la moral ha dado esa contestación. Siguiendo con el ejemplo, la ética de santo Tomás, que sigue la moral cristiana, responde que no puedes robar «porque así lo dice la Ley Divina»; mientras que la ética utilitarista, que obviamente sigue la moral utilitarista, responde que en ocasiones puedes robar «porque lo que está bien o está mal depende del nivel de felicidad que cause a la mayoría» —por eso, si robas para dárselo a los pobres, a estilo Robin Hood, habrá una mayoría significativa de gente feliz—.

En resumen, la moral responde a la pregunta: ¿de qué forma debo comportarme?; mientras que la ética me explica: ¿por qué debo comportarme de dicha forma?

Como la moral responde a qué debo hacer, la moral se constituye por un conjunto particular de principios, preceptos, mandamientos o prohibiciones. En palabras del cantautor Joan Manuel Serrat, la moral es la que dice «niño, deja ya de joder con la pelota, que eso no se dice, que eso no se hace, que eso no se toca»[74]. La ética por su lado reflexionaría sobre esos preceptos e intentaría dar respuesta a por qué el niño debe dejar de joder con la pelota, debe dejar de decir eso o hacer eso o tocar eso.

Atendiendo a estos dos tipos de preguntas, ya podemos tener una primera definición sobre la ética y la moral.

La ética es la rama de la filosofía que se ocupa del hecho moral. Dentro de la filosofía existen muchas ramas de conocimiento, cada una con un nombre propio. Si queremos estudiar el conocimiento, y saber, entre otras cosas, por qué conocemos o cómo sabemos si algo es verdad, tenemos una rama de la filosofía que se llama gnoseología o epistemología si hablamos del conocimiento científico —como ves, suelen tener nombres raros—. Si

queremos estudiar las cosas en general y saber qué es la materia o qué es real, acudimos a la rama de la filosofía que se llama ontología. Pues bien, de igual forma, existe una rama de la filosofía, llamada ética, que explica cuestiones sobre el hecho moral.

Fíjate que en la definición anterior hay un pequeño cambio. He dicho «el hecho moral» en lugar de «la moral», que era el término que usábamos al hablar de los dos tipos de preguntas. Son conceptos distintos, pero relacionados, como vemos en esta definición[75]:

> El hecho moral es una dimensión de la vida humana compartida por todos, que consiste en la necesidad inevitable de tomar decisiones y llevar a cabo acciones de las que tenemos que responder ante nosotros mismos y ante los demás, para lo cual buscamos orientación en valores, principios y preceptos.

Ese conjunto de valores, principios, preceptos es lo que llamamos la moral. Por tanto, el hecho moral es una dimensión humana, dentro de la cual se encuentra la moral. Esto puede parecer un juego de palabras o ganas de liarla al más puro estilo filosófico, sin embargo, tiene su importancia.

El hecho moral es una condición que tenemos los seres humanos, en virtud de la cual nos sentimos obligados a responder, ante nosotros y ante la sociedad, sobre lo que hacemos. En la historia del mito de Prometeo vimos que Zeus repartió dos dones: el don de la justicia y el don del honor. Ese don del honor es ese sentimiento de vergüenza que podemos tener si no hacemos eso que pensamos es correcto. Es decir, ese don del honor es lo que podemos entender como el hecho moral. Esa necesidad de explicar por qué hemos hecho algo. Este hecho moral, es un don, una cualidad, que solo tenemos los seres humanos. Por eso decimos que los seres humanos somos agentes morales.

Ahora bien, como resulta que, por necesidad, me siento en la obligación de responder de mis actos, porque soy un agente moral, necesito un libro de instrucciones, una lista FAQ, que me dé normas o indicaciones de qué es lo correcto y cómo debo comportarme. Ese libro de instrucciones es la moral. No existe un único libro de instrucciones, sino distintos libros, es decir, no existe una moral, sino distintas morales que me dicen cómo actuar —de ahí tanto libro aventurero de autoayuda—.

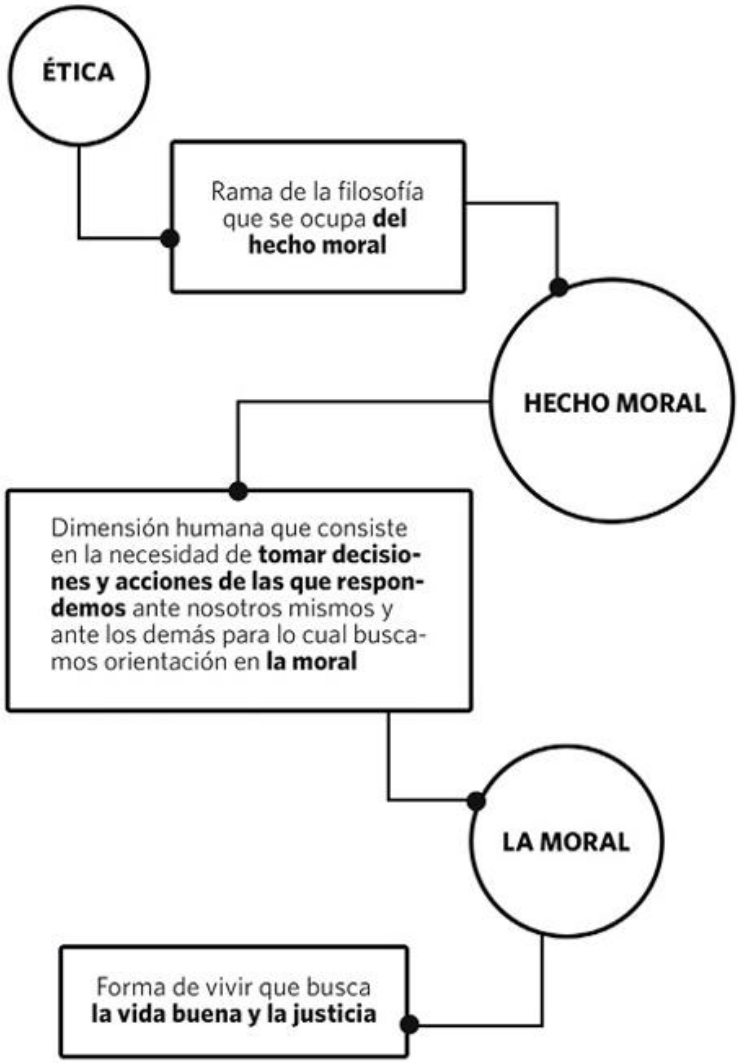
Toda moral dirá muchas cosas, pero básicamente se ocupa de dos temas, que son las preocupaciones centrales del ser humano: la vida buena y la justicia.

La vida buena es ese tipo de vida «adecuada», «correcta», con la cual nos sentimos bien. No es fácil definir cuál es esa vida adecuada o correcta y, además, depende de cada uno. Hoy en día, que rechazamos el término correcto por verlo demasiado rígido y excluyente, hablamos de vida sana, vida positiva o de comportamientos no tóxicos, términos más suaves, asépticos y menos comprometidos. Cuestión de modas. Si buscamos una definición meditada y, por tanto, más permanente de qué es esa vida buena vemos que consiste en ese tipo de vida que da sentido al destino de la persona, a su razón

de ser[76]. En principio, cada moral, si está bien construida, debe ayudarte en esa búsqueda de la vida buena —cosa distinta de la panoplia de consejos vacíos *instagramerianos* que buscan más bien la buena vida fácil—.

Por otro lado, nuestra vida siempre se realiza en sociedad por lo que también nos preocupa cómo buscar esa vida buena dentro la sociedad en la que vivimos. También es complicado, pues implica buscar un equilibrio entre la vida en común, a la vez que preservamos nuestra autonomía como individuos. Este equilibrio entre lo común y la autonomía es el principio de la justicia.

Por tanto, cada moral intenta ofrecer comportamientos que resuelvan dos cuestiones de nuestra vida: ¿qué es la vida buena?, ¿qué es la justicia? En la figura siguiente te muestro un esquema que resume todos estos puntos sobre la ética, el hecho moral y la moral, los cuales se encuentran encadenados.




***Ya vamos viendo la cuestión***

Hemos visto que nosotros, los seres humanos, somos agentes morales porque tenemos una necesidad innata de responder por lo que hacemos. ¿Y los robots o la inteligencia artificial? ¿Tiene la inteligencia artificial también esa necesidad de responder por lo que hace? Si tiene esa necesidad, ¿en qué moral busca orientación?

En el capítulo «Esto no es una pipa» veíamos que la inteligencia artificial se

diferencia de nosotros en la capacidad de pensar, es decir, en comprender nuestro entorno para dar sentido a lo que nos rodea y poder emitir un juicio. Esta capacidad de pensar, de emitir un juicio, es la que nos permite determinar si juzgamos que lo que hacemos es correcto o no. Es la capacidad que nos permite ser agentes morales. Mientras la inteligencia artificial no tenga esa capacidad de pensar, no podrá ser un agente moral. No tendrá esa necesidad de responder por lo que hace.

La inteligencia artificial no siente la necesidad de ser ética. Nosotros, sí. Nosotros, como agentes morales que somos, tenemos la necesidad de responder por las acciones que hacemos, en este caso, a través de la inteligencia artificial. Si usamos la inteligencia artificial, nuestra preocupación debe ser hacer lo que consideramos correcto mediante la inteligencia artificial. No es tanto que la inteligencia artificial sea ética, sino que nosotros seamos éticos a través de la inteligencia artificial.

Para ello, para ser éticos con la inteligencia artificial, buscamos orientación en la moral. Pero la moral, como hemos visto, no es única, sino que hay varias. Ahora entendemos el debate sin fin del Comité de Dirección de la *startup* AlegrIA.

Vimos que cada miembro de AlegrIA tenía una opinión fundamentada en una ética particular, posiblemente, sin que ellos lo supieran: Ariadna era algo aristotélica, Humberto seguía a Hume, Benito era utilitarista, como Bentham; Tomás hacía honor a su nombre, siguiendo a santo Tomás; Kanza era kantiana, Habana se inspiraba en Habermas y Nieves, escéptica, como Nietzsche. En los siguientes apartados vamos a ver qué significaría aplicar estas teorías éticas a un sistema inteligente, que es lo que pretende cada miembro del Comité de Dirección de AlegrIA. En particular, lo vamos a hacer para el algoritmo de inteligencia artificial que recomienda nuevos contenidos a los usuarios de la *startup*. Es decir, veremos cómo sería un sistema inteligente de recomendación de contenidos que fuera, por ejemplo, aristotélico, kantiano, utilitarista o nietzscheano.

## Porque buscamos la felicidad

### *Aristóteles: Lo que le haga feliz*

«¡Pues lo que haga más feliz al usuario! Todos buscamos la felicidad y tenemos que transmitir la idea de que esa felicidad es posible. ¡Con esa visión nació nuestra empresa!».

Para Aristóteles[77], debemos comenzar por el fin y, en particular, por el fin particular de cada actividad. Cada actividad tiene un objetivo, un fin determinado. Por ejemplo, el fin de la medicina es la salud, el fin de la estrategia es la victoria o el fin de la arquitectura es un edificio. Por ello, si quiero actuar de forma correcta en una actividad, primero debo saber cuál es el fin de dicha actividad. A ese fin Aristóteles lo llama «el bien» de esa actividad. Así, el bien de la medicina —lo que está bien en la medicina— es la salud. No existe «el Bien» genérico, sino «un bien» para cada cosa. Dicho en lenguaje más filosófico, para Aristóteles el bien de cada cosa es su fin.

De manera general, como ser humano, si quiero hacer lo correcto, lo primero que tendré que pensar es cuál es mi objetivo, mi fin como persona. Pero, ¿cuál es el fin del ser humano?, ¿cuál es mi fin en esta vida? Gran pregunta que, posiblemente, nos hemos hecho en alguna ocasión. Aristóteles lo resuelve rápidamente: el fin del ser humano es la felicidad. Nuestro objetivo en esta vida es ser felices. Por tanto, lo que está bien para nosotros es la felicidad. Actualmente, Aristóteles parece estar de moda, si atendemos a tanto discurso *happy flower* que escuchamos. Pero no tanto.

Porque, ¡cuidado!, no vale cualquier tipo de felicidad. Aristóteles habla de una felicidad algo especial, de una «felicidad verdadera» llamada *eudaimonía*. Podemos pensar que la felicidad es tener riquezas, éxito o millones de seguidores en Instagram. Pero Aristóteles dice que no, porque esa felicidad no te lleva al verdadero fin del ser humano. La verdadera felicidad (*eudaimonía*) es aquella que permite realizar la función del ser humano.

Ahora sí que la hemos liado, porque esto parece el típico barullo filosófico, que después de leerlo te quedas como estás. Resulta que el fin del ser humano es una felicidad verdadera, que solo sabré cuál es cuando sepa cuál es el fin del ser humano. ¡Fantástico, Aristóteles!

Para salir de este lío, Aristóteles define cuál es la función del ser humano: «La función del ser humano es el ejercicio de las actividades del alma de acuerdo con la virtud y durante una vida entera». Responde, pero ahora introduce tres elementos: alma, virtud y vida entera, que tenemos que ir desgranando.

Cuando Aristóteles habla del ejercicio de las actividades del alma, se refiere, en realidad, al ejercicio de la razón en dos dimensiones: en su parte más intelectual o cuando controlamos nuestras apetencias o pasiones de forma razonada. Para Aristóteles el fin del hombre es ejercitar su razón en estos dos

ámbitos.

Pero no de una forma cualquiera, sino de forma excelente, que es la virtud (el segundo elemento de la definición). La virtud es el modo especial de ser de cada cosa, es la excelencia de su función. Así, por ejemplo, la virtud del cuchillo es cortar o la virtud del ojo es ver. Y, ¿cuál es la virtud en el ser humano?, ¿cómo realiza el ser humano la excelencia en su función? Aristóteles responde: «La virtud es la disposición o hábito de elegir el término medio relativo a cada uno en acciones y emociones, guiado por la razón y como lo haría una persona prudente». Esta definición no tiene desperdicio y dice muchas cosas:

- La virtud es una *disposición y es un hábito*. Hay que querer ser virtuoso y esto solo se consigue con la práctica.
- Es una disposición que te lleva a *elegir*, a seleccionar, el *término medio* en las acciones y emociones. Por ejemplo, si queremos ser valientes, deberemos escoger el punto medio entre el miedo y la audacia; si queremos ser generosos, debemos elegir el término medio entre el despilfarro y la tacañería.
- Además, es un término medio *relativo a cada uno*, es decir, que no es único, sino que depende de cada persona. Cada individuo es distinto y necesita conocer su punto medio. Esto solo se consigue, de nuevo, con la práctica.
- La persona virtuosa ejercita la razón y elige con *prudencia*. Esta prudencia nos permite escoger lo más oportuno en cada momento respecto a lo que es bueno o malo para nosotros. Para ser prudente hay que deliberar, contrastar opiniones, hay que entender bien la situación y saber juzgar.

La virtud, en resumen, es esa disposición que nos permite elegir con *moderación* (término medio) y *prudencia*. Moderación y prudencia, dos elementos relevantes que después usaremos.

Todo esto de la virtud venía por la definición que vimos de la función del ser humano, que es el ejercicio de las actividades del alma de acuerdo con la virtud y durante una vida entera. Hemos visto a qué se refiere Aristóteles con las actividades del alma (ejercitar la razón y controlar las pasiones) y ahora con la virtud (elegir con moderación y prudencia). Queda la cuestión de la vida entera.

Esto Aristóteles lo resuelve de una manera muy poética diciendo: «una golondrina no hace verano». Ser virtuoso, realizar de forma excelente la función del ser humano no es cosa de una vez, sino de una vida entera. No basta que una vez elijas bien, con moderación y prudencia. Eso es una golondrina casual en el cielo. Ser virtuoso es cuestión de práctica, de probar y errar, y de elegir después de todo de forma consistente con moderación y prudencia. Eso son cientos de golondrinas, que avisan de la verdadera presencia

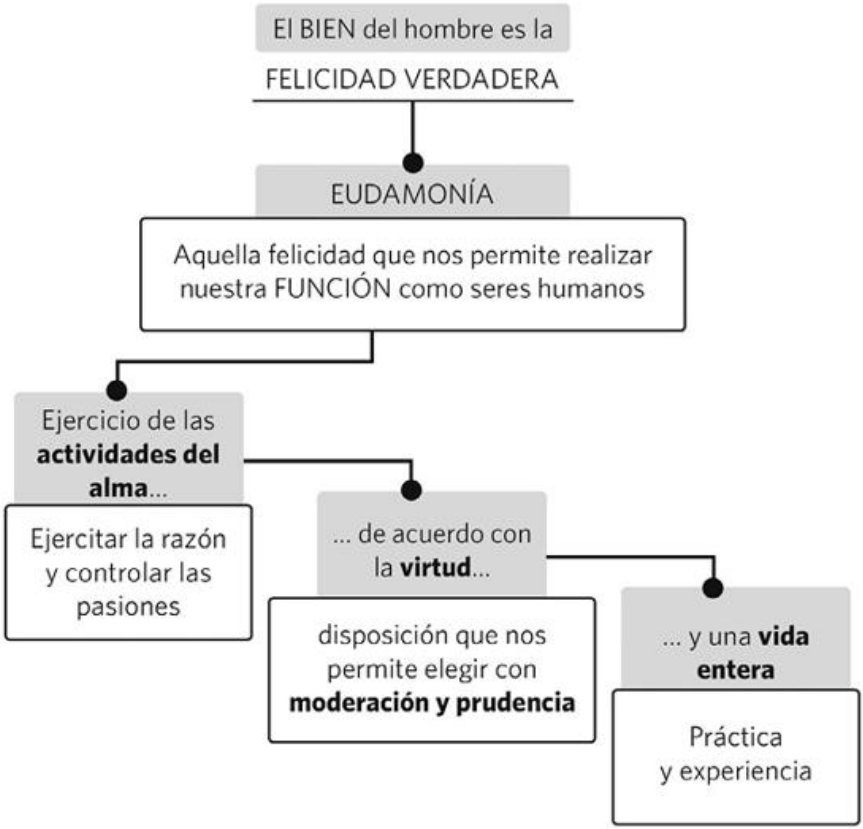
del verano. Aparece, una vez más, esa visión de la experiencia, de la práctica. Ser virtuoso, elegir el punto medio, con moderación y prudencia, no lo vas a aprender viendo *reels* de Instagram. Hay que ponerse manos a la obra, actuar y practicar. ¡Vale!, pero, ¿sobre qué actúo?

Aristóteles propone una serie de virtudes, es decir, una serie de acciones y emociones sobre las cuales encontrar ese punto medio. Así, habla del valor, la templanza, la generosidad, la gentileza (punto medio entre la ira y el servilismo) o la magnanimidad, también llamada grandeza de alma. Esta última es una virtud curiosa porque es reconocerse uno mismo merecedor de grandes cosas, siendo el punto medio entre ser vanidoso y pusilánime. El que tiene grandeza de alma, según Aristóteles, se comporta de la siguiente manera: no pide favores y, si pide algo, lo hace con desgana, mostrando poco interés; camina lentamente y habla con voz profunda. La grandeza de alma es opuesta a la humildad, la cual se considera un vicio.

Vemos que las virtudes que propone Aristóteles no con como las nuestras. Representan más bien valores propios de la Grecia antigua, que son solo para ciudadanos libres, adultos y acomodados. No hay virtudes para esclavos, trabajadores, artesanos o para las mujeres. Sobre las mujeres, Aristóteles indica que reciben el gobierno de sus maridos y atienden lo que estos les encomiendan y según la división de tareas que tiene cada uno.

Por último, un dato curioso. Aristóteles introduce un elemento de fortuna para ser feliz. Así, por ejemplo, si no tienes la suerte de haber nacido en el sitio adecuado o de tener belleza, es más complicado que seas feliz. Aristóteles habría arrasado hoy en día con esto último, donde se confunde lo bueno con lo bello.

Una vez vistos todos los elementos, corresponde hacer una breve recapitulación, que además te muestro en un gráfico. Según Aristóteles, el fin (el bien) del ser humano es la felicidad. Pero no una felicidad cualquiera, sino una felicidad verdadera (*eudaimonía*), que es aquella que nos permite realizar nuestra función. Esta función es el ejercicio del alma (ejercitar la razón y controlar las pasiones) por medio de la virtud y durante toda tu vida. La virtud es un modo de ser que te permite elegir ese punto medio (moderación) con prudencia (deliberando y sopesando). Esta felicidad verdadera que propone Aristóteles es en definitiva un tipo especial de carácter que solo se consigue con la práctica y el esfuerzo personal.




***La felicidad aristotélica en la inteligencia artificial***

Ariadna defendía que había que ofrecer a cada usuario los contenidos que les hicieran más feliz. En un principio, Aristóteles estaría de acuerdo con ella, puesto que el bien del ser humano es la felicidad. Cumplir esta propuesta parece inicialmente sencillo. Necesitaríamos algún tipo de medición de la felicidad del usuario, que bien pueden ser las estrellas que asigne a un contenido, o la frecuencia con la que ve un cierto tipo de contenidos, entendiendo que el usuario ve con más frecuencia aquellos contenidos que más feliz le hacen. La inteligencia artificial es muy buena cumpliendo objetivos. Por ello, si le pedimos que recomiende contenidos para que cada vez el usuario ponga más estrellas o esté más tiempo en la plataforma, seguro que lo consigue. Aristotélico total. Pero, ¿estar conectado es estar feliz?, ¿qué pasa con la moderación y la prudencia?

Aristóteles nos habla de una felicidad un tanto especial, de una felicidad verdadera (*eudaimonía*) a través de la cual el ser humano realiza su función, que

es el ejercicio de la razón de forma virtuosa. Esta virtud se alcanza mediante el hábito de elegir el punto medio y como lo haría una persona prudente. Si de verdad queremos ser aristotélicos, tenemos que considerar la cuestión de la virtud, a través de la moderación y la prudencia. ¿Habrá pensado Ariadna en este punto de incluir la moderación y la prudencia?

De hacerlo, de querer ser aristotélicos, en ocasiones habría que ofrecerle al usuario contenidos quizás no del todo de su gusto, para que pudiera tener un punto medio. ¿Se atrevería AlegrIA a ofrecer contenidos que sabe que no le van a gustar al usuario?

La prudencia es incluso más complicada de cumplir. AlegrIA tendría que ofrecer información detallada sobre los pros y contras de un contenido. Cosa difícil, no desde un punto de vista técnico, sino porque, ¿quién informa hoy en día de los posibles perjuicios o fallos de un contenido? ¿Estaría dispuesta AlegrIA a decir algo crítico sobre un contenido? ¿Qué le diría a la productora de tal contenido: «es que somos aristotélicos»?

En principio el comentario de Ariadna parece aristotélico, y es de gran actualidad, pues hoy se alaba la felicidad. Pero es una felicidad en forma de *likes*, seguidores, o estrellas de valoración; sin moderación ni prudencia. Porque incorporar esto es más complicado de programar y molesta al usuario. Si de verdad queremos una ética de Aristóteles, tenemos que ejercitar la virtud, mediante la moderación y la prudencia. Y por una vida entera, que una golondrina no hace verano, ni ver un documental de La 2 un día no te hace intelectual.

Pero supongamos que Ariadna es completamente aristotélica y defiende una recomendación de contenidos que lleven al usuario a esa felicidad verdadera (*eudoimonía*). Supongamos que programan su sistema con los elementos de moderación y prudencia. El siguiente punto sería: ¿qué virtudes busco en los usuarios? En un principio tendría virtudes aristotélicas, tales como el valor, la templanza, la generosidad y la grandeza de alma. ¿Son estas las únicas virtudes? ¿Y si el usuario quiere otras? Además, Aristóteles solo pensó en virtudes para ciudadanos libres varones, pues las mujeres estaban al gobierno de sus maridos. ¿Propondrá AlegrIA contenidos para esta visión? Y, ¡ojo con los feos!, que, según Aristóteles, lo tienen más difícil para ser felices; quizás no merezca la pena ofrecerles contenidos. Ser aristotélico ya no está tan bien.

Una última cuestión. Aristóteles dice que el fin del ser humano es la felicidad, pero esto no lo demuestra, lo da por hecho y por comúnmente aceptado. ¿Por qué la felicidad? Porque es lo normal, lo natural, lo que dice el sentido común, se podría contestar. Algo parecido argumentaba Tomás en la discusión del Comité de Dirección de AlegrIA. Vamos a ver este tema de *lo natural*, el sentido común.

### *Santo Tomás: lo natural*

—Bueno, ¡vamos a ver! Esto tampoco puede ser lo que quiera cada uno o lo que quiera la mayoría. ¡Así nos va! —saltó Tomás, con su gesto de complacencia habitual siempre que hacía una crítica a la sociedad actual—. De alguna manera todos sabemos que lo que está bien o no. Apliquemos el sentido común, ¡por favor!

Con la llegada del cristianismo entran en el ámbito de la ética elementos tomados de la Biblia y de la revelación de Jesucristo. La preocupación de los grandes teólogos cristianos consiste en razonar los principios morales que establece el cristianismo, para, de esta forma, crear una doctrina coherente y sólida que salga del ámbito de la fe para entrar en el orden de la razón.

Uno de tales teólogos es santo Tomás de Aquino, autor de la *Suma Teológica*[78]. Habitualmente aparece representado con un libro, o escribiendo su doctrina, e iluminado por la gracia divina. En su pecho brilla un sol que representa su sabiduría.

Santo Tomás parte de la ética aristotélica y la adapta al cristianismo. Coincide con Aristóteles en que el fin del hombre es la felicidad, pero ahora esa felicidad verdadera no es la *eudaimonía*, sino lo que llama la bienaventuranza, que es la contemplación de Dios en su esencia. También coincide con Aristóteles en que esta felicidad solo se consigue a través de la virtud. Sin embargo, santo Tomás difiere de Aristóteles respecto al origen de las virtudes y en aportar nuevas virtudes.

Para Aristóteles la virtud es un logro de nuestra razón. Es una capacidad que tenemos en nuestra naturaleza como seres humanos. Para santo Tomás esa capacidad de ser virtuoso ha sido infundida en nosotros por Dios sin intervención nuestra, a través de la ley natural y la ley divina. ¡Cuántos conceptos en un párrafo tan corto! Virtud, ley natural y ley divina. Vamos a verlos.

La bienaventuranza se consigue mediante una vida correcta. Necesitamos entonces saber qué es correcto, es decir, qué está bien y qué está mal. Esto lo sabemos gracias a la ley natural que Dios ha inscrito en nuestro ser. La ley natural es una ley común a todo ser humano, que nos inclina de forma natural hacia el bien y nos permite discernir lo bueno de lo malo a través de la razón. Gracias a esta ley natural podemos intuir qué está mal o qué está bien, y por ello podemos ser virtuosos. Por eso, cuando algo está mal, es que está mal, no hay vuelta de hoja. ¡Es natural!

Ahora bien, ¿por qué tenemos la ley natural que tenemos? Es decir, según esta ley natural, entendemos que está mal matar o robar. ¡Es natural que esté mal! ¿Por qué? Porque esta ley natural forma parte de la ley divina. La ley divina es la intención de Dios, es lo que Dios quiere. Si nuestro fin es la bienaventuranza, que es la contemplación de Dios en su esencia, debemos tener la orientación de hacia dónde mirar. Esa orientación es la ley divina, que no podemos conocer, pero sí podemos intuir a través de la ley natural que

tenemos inscrita.

Esta existencia de la ley natural y su correspondencia con la ley divina es lógica. No tendría sentido que la ley divina y la ley natural fueran contrarias. No sería razonable que, para alcanzar esa bienaventuranza, Dios nos pidiera algo (ley divina) que no pudiéramos realizar (que no estuviera en nuestra inclinación natural). Por ello la ley divina, lo que Dios nos pide que hagamos, es algo que podemos hacer de forma natural gracias a esa ley natural con la que nos ha dotado. No tendría sentido que Dios me pidiera no matar, pero yo no entendiera por qué. Ahora lo divino, lo natural y lo racional: lo que Dios me pide, es algo natural en mí y lo entiendo. ¡Natural!

Luego, efectivamente, para Aristóteles podemos ser virtuosos si ejercitamos la razón, como lo haría una persona prudente, y podemos discernir por nosotros mismos qué está bien o qué está mal. Para santo Tomás, esto lo sabemos, no por nosotros mismos, sino por la ley natural que nos ha infundido Dios.

También dije que santo Tomás incorporaba nuevas virtudes respecto a Aristóteles, las llamadas virtudes teologales: fe, esperanza y caridad. Solo mediante estas virtudes, también infusas en nosotros, podemos llegar a alcanzar esa bienaventuranza de contemplar a Dios en su esencia. Las virtudes aristotélicas están muy bien, y debemos seguirlas, pero con ellas tan solo podemos alcanzar un simulacro de bienaventuranza, ser algo felices y sentir un poquito a Dios. Las buenas de verdad, las virtudes *Premium*, son las virtudes teologales, porque solo ellas me llevan a la bienaventuranza plena.

### *Programemos lo más natural en la inteligencia artificial*

Santo Tomás no difiere tanto de la ética aristotélica, salvo que incluye a Dios, como fin del ser humano, la ley natural, inspirada en nosotros por la ley Divina, y las virtudes teologales, que nos permiten alcanzar esa bienaventuranza. Por tanto, las objeciones que pusimos para aplicar la ética de Aristóteles a la inteligencia artificial siguen siendo válidas para el caso de santo Tomás. Ahora bien, estos nuevos elementos tienen su importancia.

Tomás —ahora me refiero al de la *startup* AlegrIA— defiende una visión natural de lo que está bien y está mal. Posiblemente él no lo sepa, pero este pensamiento está inspirado en esa idea de una ley natural inscrita en nosotros. Debido a esta ley natural, las virtudes son algo razonable. Es lo natural. Si queremos aplicar la ética de santo Tomás al sistema inteligente de recomendación de contenidos de AlegrIA, no tendríamos que dar muchas explicaciones de por qué el sistema recomienda algo. Es lo natural. Sería natural que se recomendaran contenidos para favorecer la fe, la esperanza o la caridad. Pero esto, que puede parecer natural, si somos tomitas, puede que no lo sea para otro y que este otro diga, con el mismo aplomo, que lo natural es otra cosa.

En el fondo, Aristóteles también defiende esta visión natural, pero de otra forma. Él dice que lo natural es que el ser humano busque la felicidad y a ese fin natural lo llama el bien. ¡Buen salto! ¿Por qué si la felicidad es nuestro fin natural, la felicidad es lo que está bien? ¿Por qué lo que supuestamente puede ser lo natural en nosotros es en definitiva lo bueno? Este pensamiento lo tenemos muy extendido actualmente, en el que se defiende que todo lo que nos pasa, si es natural, es entonces bueno. ¡Deja fluir tus emociones, que es lo natural, y eso es bueno!

Las emociones de cada uno, justo lo que decía David Hume, nuestro próximo invitado a esta reunión de filósofos. Porque tras la llegada del Renacimiento pasamos a pensar en éticas que consideren la individualidad y la conciencia de cada uno.

## Según cada uno

### *Hume: Lo que más le guste*

—Vale, pero qué tipo de felicidad, porque esto depende de los gustos de cada uno —replicó Humberto con autosuficiencia.

David Hume es lo que se llama un filósofo empirista, que se opone a los llamados filósofos racionalistas. También en la filosofía, como en el futbol, hay rivalidades. Los empiristas defienden que el ser humano conoce las cosas a través de la experiencia práctica y, en particular, gracias a los sentidos. Sabemos de las cosas, no porque primero las pensemos y luego las experimentemos, sino porque primero las experimentamos y luego las pensamos —algunos incluso nunca las piensan—. Un empirista diría que aprendemos en la universidad de la vida. Para un empirista lo importante son los sentidos.

Desde esta perspectiva basada en los sentidos, Hume establece de manera clara que la razón no tiene nada que decir sobre la moral[79]. No merece la pena que nos pongamos a razonar por qué algo está bien o por qué está mal. Intentar encontrar una explicación lógica a por qué no hay que robar es perder el tiempo. Según Hume, la razón nunca podrá averiguarlo.

Dicho de una manera más precisa, en lenguaje filosófico, los juicios morales no son racionales. Un juicio moral es aquel que dice, «esto está mal» o «esto está bien», y eso nunca lo podremos deducir pensando. La razón nos ayuda a entender un hecho en lo que se refiere a qué ha sucedido. Por ejemplo, en un robo, entendemos que hay un objeto que pertenece a una persona, que ahora lo tiene otra en su poder y en contra de la voluntad de su dueño. Esto se llama describir un hecho. La razón emite juicios descriptivos —también se pueden llamar juicios fácticos o juicios de hechos—. Es lo mismo que escuchamos en las sentencias cuando se dice «queda probado». En el caso de un robo diríamos: «queda probado que el demandante es el dueño del teléfono móvil y queda probado que el demandado se los sustrajo». Hasta ahora tan solo hemos descrito los hechos, pero no hemos valorado la acción. Hemos hecho un juicio descriptivo, y falta un juicio de valor.

Este juicio de valor es lo que hace la moral y consiste en aprobar o desaprobar un hecho. Una vez descrito el hecho del robo, me queda por emitir un juicio de valor, que dirá si apruebo o no tal robo. Tras realizar el juicio descriptivo, realizo el juicio de valor. Para Hume, entre ambos juicios no hay ninguna conexión racional. Si yo emito el juicio de valor «está mal robar», no hay ninguna razón que lo explique. Todo es cuestión de las pasiones y los sentimientos.

Las cosas funcionan de la siguiente manera. Primero tengo una pasión, que me hace actuar de una cierta forma, luego la razón describe ese hecho y, posteriormente, los sentimientos juzgan el hecho. Una pasión me lleva a robar,

la razón describe el hecho de un teléfono móvil que ya no lo tiene su dueño, y finalmente el sentimiento me dirá «eso está bien», o «eso está mal», dependerá de los sentimientos de cada uno.

Para Hume, la moral se siente, no se razona. Uno siente que algo está bien o que algo está mal. No lo razona, lo siente. Si uno dice «robar está mal», lo que en realidad está diciendo es «a mí me parece mal robar, no me gusta robar». Es una cuestión de sentimientos. Hay una frase de Hume, muchas veces repetida, que dice: «no es contrario a la razón preferir la destrucción del mundo a herirme en un dedo». La razón no tiene nada que decir respecto a si es mejor destruir el mundo o cortarme un dedo, solo describirá la destrucción del mundo o el corte en mi dedo. Otra famosa perla de Hume: «la razón es y debe ser esclava de las pasiones y no puede pretender otra cosa sino servirlas y obedecerlas». Quizás no nos demos cuenta, pero hoy en día hay mucho empirista suelto. Basta escuchar frases como «haz lo que te pida el cuerpo, lo que te diga tu corazón»; o bien, «si eso es lo que siente, está bien». Hoy Hume tendría más *followers* que Rosalía.

Esta falta de conexión entre la razón y la moral, entre describir un hecho y juzgarlo, tiene una consecuencia que sacaba de quicio a Hume. Es el hecho de extraer un «deber ser» a partir de un «ser». Lo vimos en el capítulo primero al hablar de la obra *Caos Primordial* de Hilma af Klint cuando hablábamos del debate entre «lo que es» y «lo que debería ser».

La razón describe un hecho, «esto es robar», mientras que la moral emite un juicio de valor que describe un sentimiento de aprobación o desaprobación, «no se debe robar», que en realidad significa «no me parece bien robar». Fíjate que la primera frase tiene el verbo «ser», mientras que la segunda tiene el verbo «debe ser». Lo indignante para Hume era que en los libros de moral durante un buen rato se hablaba de lo que era y de repente se pasaba a lo que debía ser, sin explicar tal salto. Se pasaba de hablar sobre robar a decir que no se debe robar, sin explicar por qué. Para Hume era claro por qué no se explicaba, porque no se puede, porque la frase «no se debe robar» quiere decir «no me gusta robar». Es una cuestión de sentimientos.

Vale, Sr. Hume, no podemos explicar de forma razonada por qué decimos «no se debe robar». Pero si esta desaprobación depende de los gustos y los sentimientos, de cada uno, ¿cómo es posible que exista un consenso entre todos los seres humanos sobre la desaprobación de robar? Por mucho que lo intentemos, Hume no se deja pillar y responde: debido a dos factores que son la utilidad y la simpatía.

La utilidad es otro sentimiento por el cual valoramos positivamente (aprobamos) aquellas acciones que nos resultan útiles, es decir, convenientes o beneficiosas. En general me dedico a aquello que resulta provechoso. La simpatía es un mecanismo por el cual convertimos los sentimientos de los demás en nuestros sentimientos. Se produce porque somos semejantes unos a

otros y así, cuando alguien comparte conmigo sus sentimientos, yo llego a hacer míos esos mismos sentimientos. Gracias a la simpatía nos alegramos por las alegrías de otros y nos apenamos por las desgracias de otros. Si unimos los dos conceptos, la simpatía hace que dirija mis esfuerzos hacia lo que es útil para los demás. Por eso no me gusta la idea de robar. Si todos nos dedicamos a robarnos unos a otros, esto no sería útil para vivir en sociedad.

Como dice Hume, la utilidad y la simpatía explican por qué valoramos la generosidad, humanidad, compasión, gratitud, amistad, fidelidad, celo, desinterés, liberalidad y todas las demás cualidades que hacen que una persona sea considerada buena: porque esto hace que sea una persona agradable y útil. Al final, llegamos a la misma conclusión y a valorar unas mismas virtudes, pero por otro camino. ¡Qué simpático este Hume!

### *Inteligencia artificial a golpe de like*

Como bien decía Humberto, en la *startup* AlegrIA lo que agrada o desagrada depende de los gustos de cada uno, que es igual a decir, que lo que está bien o mal depende de los gustos de cada uno. Afortunadamente, gracias a la simpatía tenemos algunos gustos comunes, gracias a la utilidad buscamos cosas provechas y gracias a ambas buscamos cosas provechosas comunes.

Si la empresa AlegrIA quiere ofrecer contenidos según la ética de Hume, estos deberán encontrar esta utilidad y simpatía. Quizás no baste con la típica escala de estrellitas o *likes* para clasificar un contenido. Si alguien pone cinco estrellas a una película, no sabríamos si es por simpatía, porque le suscitó emociones similares o por utilidad, porque le sirve para la vida. Habría que pedir al espectador dos calificaciones: una por simpatía y otra por utilidad.

Posteriormente, AlegrIA ofrecería aquellos contenidos con mayores índices de simpatía y utilidad, pues cabe entender que aquello más simpático y útil para la mayoría, será lo más agradable y beneficioso para alguien en particular. Llegaríamos así a la tiranía de la mayoría, en defensa de grandes sentimientos.

No bastaría con ofrecer aquellos contenidos que le gustaran a cada uno de manera particular, pues quizás no serían útiles o no causarían simpatía en el resto de la audiencia. Si a uno le gustan los documentales sobre la vida de los protones y otras partículas subatómicas, estaría fastidiado, pues posiblemente este tipo de documental no tendría los mayores índices de utilidad y simpatía —y pido perdón por si existe algún colectivo en defensa de los protones, y se sienten ofendidos; mi simpatía hacia ellos—.

O bien, si la mayoría de la audiencia tiene por oficio abrir butrones para robar, AlegrIA solo ofrecería películas de ladrones y documentales de cómo reventar cajas fuertes. Esto sería lo más útil y simpático para la mayoría de la audiencia.

¿Es esto razonable? Hume respondería que por qué hablamos de razonable o no, cuando la razón no tiene nada que decir al respecto. Son los sentimientos

de utilidad y simpatía, nada de pensamiento. Ya… Pero vuelvo a preguntar, sin que me oiga Hume, ¿es esto razonable?

Kant dice que no ¡Categóricamente que no!

### *Kant: Porque sí y punto*

—Pues partamos entonces de unos principios. ¡Tiene que haber calidad! —replicó con energía Kanza.

En alguna ocasión, cuando éramos pequeños y pedíamos razones sobre por qué teníamos que hacer algo, la respuesta de nuestros padres era «porque sí y punto». Posiblemente esa contestación era consecuencia de su hastío por nuestra perseverancia en no hacer ese algo, pero desde un punto de vista filosófico, es pura visión de Kant.

Como hemos visto, y a modo de resumen *twitteriano*, para Aristóteles hacemos el bien porque nos da la felicidad (*eudaimonía*), para santo Tomás, porque nos da la bienaventuranza y para Hume, por utilidad y guiados por la simpatía. En todos estos casos hacemos algo que está bien para conseguir un fin: la felicidad, la bienaventuranza o la utilidad. Hacemos el bien por un bien mayor, porque obtenemos un beneficio.

Hay personas que esta visión les parece pobre moralmente hablando. Cuántas veces hemos escuchado algo parecido a: «sí, le ayudó para quedar bien», o «estuvo con él porque le convenía». Parece que hacer el bien pierde mérito si con ello conseguimos algo. La bondad se convierte entonces en una transacción, donde uno ofrece una buena acción y a cambio recibe un beneficio. En esta visión, la moral buena, la buena, buena de verdad es la que hace el bien porque sí, sin esperar nada a cambio. Dicho en lenguaje filosófico, la moral buena es la que se mueve por imperativos categóricos, en lugar de imperativos hipotéticos.

Un imperativo hipotético es aquel que te obliga a hacer algo bajo una hipótesis, es decir, partiendo de una condición. Lo podemos expresar de la forma: «Si quieres algo, entonces debes hacer esto». Sin embargo, en un imperativo categórico no hay condicionantes. El imperativo categórico te obliga a hacer algo porque sí. No se expresa partiendo de ninguna condición, sino que simplemente dice: «Debes hacer esto». Este tipo de imperativo categórico es el que le gustaba a Kant. Vamos a ver por qué.

Según Kant[80], lo único bueno es la *buena voluntad* —ya empezamos con los juegos de palabras—. Quiere decir que algo es bueno, no por el resultado que causa, por ejemplo, ayudar a otro, sino que es bueno si se hace desde el principio con la voluntad de querer hacer algo bueno. Lo bueno es actuar *por deber* y no *conforme al deber*. Actuar por hacer algo bueno y no hacer algo bueno de casualidad, buscando otra cosa. Si haces donaciones benéficas porque desgravan, estás actuando conforme al deber de ayudar, pero no por el deber de ayudar. Lo bueno es la buena voluntad, que es aquella que actúa por deber.

Vale, pero, ¿qué deber? El deber moral. ¿Y qué es el deber moral?

Este deber es una ley y, como tal, es exclusiva de los seres humanos. Solo los seres humanos tenemos la capacidad de crear leyes, que nacen de nuestra razón, y por las cuales ordenamos la realidad. Entre esas leyes se encuentran unas leyes particulares llamadas leyes morales. Por tanto, la buena voluntad es aquella que actúa por deber, que cumple porque sí con una ley moral. Como esta ley moral es ley y es porque sí, se le llama entonces un imperativo categórico.

Unimos ahora todas las piezas: lo que está bien es la buena voluntad, que es aquella que actúa bajo una ley moral. Lo que está bien es actuar bajo un imperativo categórico.

El siguiente paso es saber si esa ley que nos creamos desde la razón es una verdadera ley moral. Porque podemos crear muchas leyes, pero no todas serán leyes morales, y, por tanto, no todas nos llevarán a la buena voluntad. Para saber si estamos actuando bajo una ley moral, Kant nos ofrece las llamadas formulaciones del imperativo categórico, que son como una prueba del nueve para saber si una ley que nos creamos es verdadera ley moral.

Según Kant, obramos de forma moral si nuestra máxima, es decir, el pensamiento que guía nuestra conducta, cumple estos tres principios:

- Universalidad: Obra solo según una máxima tal, que puedas querer al mismo tiempo que se torne en ley universal.

Toda ley que se precie es universal. Por ejemplo, la ley de Newton es universal, aplica a todo el Universo. Una ley de Newton que dijera que en unos sitios los cuerpos caen hacia abajo, y en otros sitios caen hacia arriba, ni es ley ni es nada. Por tanto, nuestras leyes morales, si de verdad son leyes, deben ser universales. Dicho de otra manera, si una supuesta ley no es universal, entonces no es ley moral. Se me puede ocurrir una ley que, por ejemplo, diga "roba los zapatos al otro", pero esta ley no la puedo convertir en universal, pues todos acabaríamos descalzos y eso no es razonable. Por tanto, esa ley se me puede ocurrir, pero no sería una ley moral porque no la puedo convertir en universal.

- Dignidad. Obra de tal modo que trates a la humanidad, tanto en tu persona como en la de cualquier otro, siempre como un fin y nunca solamente como un medio.

Una vez que sabemos que la ley moral debe ser universal, el siguiente paso es dar contenido a esa ley, saber de qué trata. Como esta ley debe ser categórica, es decir, no deber ser para buscar otra cosa, el objeto de esa ley debe ser la propia humanidad. No podemos buscar otra cosa que nuestra humanidad. Una ley es moral si por encima de todo respeta la dignidad del ser humano y de la propia humanidad. Esta dignidad se consigue con dos condiciones. Una, la que indica este principio: cuando nos tratamos a nosotros mismos y al otro como

un fin en sí mismo y no como un medio para algo. La segunda condición es el tercer principio que viene ahora.

- Autonomía: Obra como si por medio de tus máximas fueras siempre un miembro legislador en un reino universal de los fines.

La segunda condición para respetar esa dignidad del ser humano es que uno mismo sea capaz de generar esas leyes morales. Es el principio de la autonomía de la voluntad. Uno es autónomo cuando uno mismo, desde la razón, decide imponerse leyes y no tiene que esperar a que otro se las imponga. Además, es autónomo de verdad si esas leyes que crea son universales y tratan a todos como un fin en sí mismos.

Después de todo esto, prepárate, porque ahora viene el colofón final. La buena voluntad es aquella que obra por deber, para respetar una ley moral, la cual debe ser universal, tratar a la humanidad como fin y nacer de nuestra autonomía. El bien es la buena voluntad, que es esa disposición permanente a realizar nuestra vida según imperativos categóricos.

Hasta ahora no lo hemos dicho, pero para que se cumpla todo esto de la buena voluntad es necesaria una condición inicial: la libertad del ser humano. Si no somos libres, no tendremos esa autonomía de la voluntad, y todo el castillo de naipes de los imperativos categóricos se nos viene abajo. Si no somos libres, no hay moral. Kant no puede demostrar esa libertad del ser humano y, por tanto, la presupone. Esperemos que sea así, de lo contrario este libro no sirve para nada.

### *Reglas categóricas en la inteligencia artificial*

En principio la ética de Kant es bastante fácil de aplicar en la inteligencia artificial. Basta con definir unos imperativos categóricos y programarlos. Como son categóricos, no hay que supeditarlos a una condición previa. Serían órdenes de un algoritmo sencillo, del estilo: `<No hacer X> <Hacer Y>`.

Kanza, el miembro de la *startup* AlegrIA lo tiene claro, su imperativo categórico sería: `<Ofrece contenidos de calidad>`. Pero si queremos ser de verdad kantianos, no basta con inventarse una ley, debemos estar seguros de que es una ley moral, es decir, que cumple con las tres formulaciones del imperativo categórico: universalidad, dignidad y autonomía.

La primera condición, universalidad, parece que la cumple, pues todo el mundo quiere contenidos de calidad. Es una ley que se puede aplicar de forma universal. El problema viene cuando queremos concretar esa idea de calidad en un contenido particular. Cada persona tiene un concepto distinto de calidad. Por eso en Youtube nos encontramos vídeos de cualquier descerebrado haciendo patochadas; seguro que su audiencia piensa que son de calidad. Luego el concepto de calidad, o bien no es tan universal, o bien resulta complicado de

concretar.

Tampoco está clara la condición de dignidad. Se supone que un imperativo categórico no busca ningún fin salvo la propia humanidad. Esto significa que AlegrIA tendría que proponer contenidos de calidad, porque sí, por el mero objetivo de ofrecer calidad, sin pensar en algo más, como obtener beneficios económicos y esas cosas mundanas. Si resultara que AlegrIA ofrece contenidos de calidad para ganar dinero, ya no sería buena voluntad, porque perseguiría un fin distinto del propio deber de ofrecer calidad.

La condición de autonomía parece que la cumple AlegrIA, pues concede a cada miembro de su Comité de Dirección la posibilidad de proponer un posible imperativo categórico, como es el caso de lo que propones Kanza. Pero si queremos ser kantianos de verdad, quizás deberíamos otorgar esa autonomía también a los clientes Cada usuario podría establecer su imperativo categórico y comunicárselo a AlegrIA, para que esta lo aplicara en la recomendación de sus contenidos, si bien antes debería verificar que cumple con los requisitos de universalidad y dignidad.

Un jaleo esto de ser kantiano.

Al margen de estas discusiones, el verdadero problema de intentar aplicar la ética kantiana es doble. Primero, que no sabemos exactamente qué debemos aplicar. La ética de Kant es lo que se llama una ética formal. Es una ética que no te dice lo que tienes que hacer, sino que te da reglas para que decidas lo que tienes que hacer. La ética de Aristóteles te dice lo que tienes que hacer: practica una serie de virtudes, porque eso te llevará a la felicidad. La ética de santo Tomás también te dice lo que tienes que hacer: practica las virtudes cardinales y teologales, porque ellas te llevarán a la bienaventuranza. Pero la ética de Kant no ofrece ninguna orden sobre haz esto o aquello. Ofrece tres reglas para que determines si lo que se te ha ocurrido hacer es ley moral.

Así resulta muy complicado aplicar Kant en la inteligencia artificial. ¿Qué leyes programo? ¿Son universales? ¿Persiguen algún objetivo? Incluso si encuentro alguna ley apropiada, surge un segundo problema, respecto a las consecuencias de su aplicación. Es el caso de, por ejemplo, las famosas Leyes de la Robótica de Isaac Asimov.

En la obra Runaround[81] el científico y divulgador Asimov cita por primera vez dichas leyes de manera unificada, que aparecieron posteriormente en su recopilatorio de relatos de ciencia ficción, *Yo Robot*[82], que es una de sus obras más famosas. Según el autor, todo robot debe cumplir las siguientes leyes en su comportamiento:

- Primera ley: Un robot no hará daño a un ser humano o, por inacción, permitirá que un ser humano sufra daño.
- Segunda ley: Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entrasen en conflicto con la

primera ley.

- Tercera ley: Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley.

Si bien las leyes nacieron para su aplicación en los robots, se podrían hacer extensibles a toda la inteligencia artificial. Se supone que estas leyes deben dotar a todo robot, o inteligencia artificial, de unas directrices morales que le hagan actual correctamente en cada situación.

Estas leyes de Asimov las podemos considerar como imperativos categóricos de Kant, y en particular la primera ley. Claramente es universal y busca la dignidad del ser humano, pues evita hacer daño a todo ser humano. Además, es categórica, pues ofrece un mandato no condicionado por nada previo. Viene a decir que un robot no debe hacer daño a un ser humano, y punto, sin ninguna condición previa.

Supongamos que somos capaces de concretar qué es no hacer daño —cuestión similar a lo que hemos visto con la idea de calidad— y que podemos programar al menos esta primera ley. El problema viene, como dije, en las consecuencias de su aplicación.

Es el caso en el que para hacer un cierto bien, hay que realizar un cierto mal. Supongamos un robot que ve cómo un ser humano, pongamos su dueño, se muere de hambre, lo cual supone un daño fragante. Supongamos que el robot, al igual que su dueño carecen de medios y la única forma de salvar a su dueño es robar una barra de pan en una panadería cercana. ¿Qué hará el robot? En principio no podría actuar, porque robar una barra de pan es causar daño a otro, en particular al dueño de la panadería. Eso va en contra de esa primera ley. No podría hacer nada, tan solo ver cómo su dueño se muere, lo cual le haría entrar en bucle, pues por inacción estaría haciendo daño a su dueño.

Esto que planteo no es una cuestión única de la primera ley de la robótica de Asimov. Es una cuestión de llevar los imperativos categóricos al límite de su cumplimiento.

¿Qué habríamos hecho nosotros?

Quizás una moral, de esas que llamamos férreas, habría también evitado robar una barra de pan, para no incumplir el imperativo categórico «no robarás», con el mismo resultado de morir de inanición. O quizás sopesaríamos las posibles consecuencias. Diríamos que vale más una vida humana que una barra de pan, y que el dueño de la panadería así lo entendería también. Pero en ese caso nos alejamos del imperativo categórico de «no robarás», pues estaríamos diciendo algo así como «puedes robar, si el robo es liviano y ayuda a salvar una vida». Es decir, estaríamos sopesando una acción moral en función de sus consecuencias. Esa es la visión de los utilitaristas.

### *El utilitarismo: la felicidad de la mayoría*

—Buscar la felicidad de cada uno es complicado. Mejor ofrezcamos lo que le gusta a la mayoría y así hay menos riesgo de equivocarnos —propuso, Benito.

Estamos gobernados por dos amos soberanos: el dolor y el placer; y solo ellos nos indican qué debemos hacer y qué haremos finalmente. El dolor y el placer nos indica qué está bien y qué está mal. Con esta premisa comienza la obra Introducción a los principios de moral y legislación[83] de Jeremy Bentham, dando así origen a la corriente filosófica llamada utilitarismo.

Según Bentham, dado el hecho, que no podemos negar ni evitar, respecto a que estamos sujetos al dolor y al placer y ellos determinan qué está bien y qué está mal, es necesario tratar esta sujeción de una forma racional. ¿Cómo? Definiendo lo que llama utilidad. Un objeto o acción tiene utilidad en la medida en la que produce beneficio, ventaja, placer, bien o, en definitiva, felicidad. Si algo te causa mucha felicidad, entonces tiene mucha utilidad. Y si encima causa felicidad, no solo a ti, sino a la comunidad, entonces tiene todavía más utilidad. Es una cuestión de aritmética: a más gente contenta, más utilidad. Si recibes muchos corazoncitos en un *like* de Instagram, Bentham dirá que aquello es extremadamente útil.

Esta visión de qué es útil nos ayuda a determinar claramente qué está bien y qué está mal. Está bien lo que tiene más utilidad, es decir, lo que causa más felicidad a la mayoría. Esto, dicho así, suena muy pobre, poco filosófico. Uno no estudia filosofía para decir cosas tan sencillas. Por ello, para rodear esta idea de profundidad, Bentham incorpora el llamado principio de utilidad. Esto es otra cosa. Cuando uno tiene un principio, quiere decir que es una persona seria. ¿Qué dice el principio de utilidad? Veamos que dice su autor:

> Por el principio de utilidad se significa aquel principio que aprueba o reprueba toda acción cualquiera, de acuerdo a la tendencia en la que parece aumentar o disminuir la felicidad del afectado. […] Una acción entonces puede ser dicha conforme al principio de utilidad cuando la tendencia que tiene a aumentar la felicidad de la comunidad es mayor que cualquier tendencia que tenga a disminuirla.

Y así es como se dice, de forma seria y filosófica, lo que ya dijimos antes: que el bien moral es la felicidad de la mayoría. Según este principio de utilidad, una persona actúa correctamente, desde el punto de vista moral, si opta por una acción que proporcione la mayor felicidad para el mayor número. Esto afecta también al legislador. Un buen legislador es aquel que dicta leyes que causan la mayor felicidad a la mayoría. ¡La de políticos utilitaristas que tenemos!

El tema es serio para Bentham, porque crea una aritmética para calcular la utilidad de una acción, que es equivalente a calcular la bondad de dicha acción. De esta forma, un legislador tiene un método cierto y racional para determinar si la ley que propone es moralmente aceptable.

La utilidad de una acción, depende de la cantidad de felicidad que cause en el primer momento (inmediata), y de la que pueda causar posteriormente

(futura). La cantidad de felicidad inmediata viene dada por estos factores: intensidad, duración, certidumbre, y cercanía. Mientras que la cantidad de felicidad futura se determina por su fecundidad y su pureza. La fecundidad mide con qué probabilidad esa acción generará después sensaciones del mismo tipo: si es un placer, si genera posteriores placeres; si es un dolor, si genera posteriores dolores. La pureza nos dice con qué probabilidad se generarán después sensaciones contrarias: si los placeres generarán dolores o viceversa.

Con estas variables, ya podemos calcular la cantidad de felicidad de una acción, que nos dirá si la acción es moralmente buena o no. Hasta Bentham nos ofrece una especie de algoritmo, que dice así:

```
Seleccionar una persona.
    Calcular su valor de placer.
        Valor inmediato: Sumar los valores de
        placer según intensidad, duración,
        certidumbre, y cercanía.
        Valor futuro: Sumar los valores de placer
        según fecundidad y pureza.
    Calcular su valor de dolor.
        Valor inmediato: Sumar los valores de
        dolor según intensidad, duración,
        certidumbre, y cercanía.
        Valor futuro: Sumar los valores de dolor
        según fecundidad y pureza.
    Hacer balance.
        A los valores de placer inmediato y futuro
        les restamos los valores de dolor
        inmediato y futuro.
        Si el balance es positivo, la acción es
        buena; en caso contrario, la acción es
        mala.
Repetir estos cálculos para todas las personas
afectadas.
    Realizar nuevo balance.
        Si el balance es positivo, la acción es
        buena; en caso contrario, la acción es
        mala.
```

Bentham pensaba en todo, y dos siglos antes de inventarse la inteligencia artificial, ya dotó a la humanidad de una forma de calcular la felicidad, que es la forma de determinar si algo está bien o mal. Con esta visión de futuro es fácil calcular el nivel ético de la inteligencia artificial.

Casi cien años después aparece otro utilitarista, John Stuart Mill, quien añade un matiz a esto de la cantidad de felicidad. Con notorio desparpajo dice que es mejor ser un hombre satisfecho que un cerdo satisfecho; es mejor ser Sócrates insatisfecho, que un loco satisfecho[84]. Introduce así la cuestión de la calidad en el placer. La frase plantea que un sabio, como Sócrates, aunque no esté del todo satisfecho con la vida, porque es consciente de que este mundo es imperfecto, en el fondo es más feliz que un insensato. ¿Por qué? Porque un sabio es capaz de gozar de placeres que tienen más valor, que son de más calidad.

Defiende que los placeres no solo se diferencian en cantidad, sino también en calidad, de tal forma que hay placerse inferiores y superiores. Los placeres inferiores son los meros corporales, y son deseados por todo tipo de animal. Sin embargo, los placeres superiores son aquellos que vienen del intelecto, de los sentimientos y de la imaginación, y son deseados por seres con facultades superiores, como los seres humanos. Pero no todos los seres humanos poseen estas facultades superiores en igual proporción y por ello, no todos los seres humanos disfrutan del mismo tipo de felicidad. En consecuencia, ahora la utilidad ya no es solo cuestión de una suma de valores de placer, sino que entra una valoración cualitativa, subjetiva. ¿Y cómo determinamos ese valor cualitativo, si no lo podemos medir? Preguntando a los que han gozado de ambos tipos de placeres, de los placeres superiores y de los inferiores.

De esta forma Mill plantea que, para valorar dos placeres, si hay uno al cual todos o casi todos aquellos que tienen experiencia de ambos tipos de placeres le dan una decidida preferencia, entonces, ese placer es más deseable. Es un sufragio universal de valoración de placeres. En realidad, solo un sufragio, pero no universal, porque solo pueden votar aquellos que tienen experiencia de ambos tipos de placeres, de los superiores y de los inferiores; es decir, los sabios del estilo Sócrates.

Es una pena porque Bentham lo había dejado muy claro con su algoritmo de cálculo de utilidad y ahora Mill, intenta dotar a la utilidad de un valor superior, pero complica la ecuación. Ya no valen todos los *likes* de todo el mundo, solo de aquellos con experiencia en placeres superiores e inferiores.

### *Cálculo del mayor deseo*

La ética utilitarista es el paraíso de la inteligencia artificial, al menos desde la perspectiva de Bentham. ¡Hasta nos dejó un algoritmo! Si la *startup* AlegrIA quiere ser utilitarista, como propone Benito, no tiene más que calcular el nivel de felicidad que arroja cada uno de sus contenidos en sus clientes. Lo puede

calcular según los criterios de intensidad, duración, certidumbre, cercanía, fecundidad y pureza. Les puede preguntar a sus usuarios que valore estos conceptos, y luego calcular la media. Aquellos contenidos con medias más altas, serán los contenidos más éticos. Esto puede resultar algo incómodo para la audiencia, pues son conceptos complicados de explicar, sobre todo la idea de fecundidad y pureza. Pero seguro que AlegrIA puede buscar medios alternativos de calcular tales criterios, según el comportamiento de sus clientes.

El problema viene si somos utilitaristas en el sentido de Mill. Ahora no es adecuado preguntar a todos los usuarios. Habría que preguntar solo a aquellos usuarios que hayan experimentado los dos tipos de placeres, los superiores (intelectuales) y los inferiores (sensuales). ¿Cómo identificamos a tales usuarios? ¿Les pedimos estudios? ¿Una relación de experiencias?

Al margen de estas ideas que dejaron Bentham y Mill, difíciles de aplicar, detrás del utilitarismo se encuentra una idea más profunda. Está bien y es moralmente aceptable, aquello que causa la mayor felicidad a la mayoría. Solo hace falta ofrecer aquellos contenidos con más *likes*. Con ello, no es que estés ofreciendo los contenidos que más gustan, sin más, es que, según el utilitarismo estás ofreciendo los contenidos más éticos.

Esto se puede trasladar a cualquier actuación de una inteligencia artificial. Volvamos al caso del robot que vimos en la ética de Kant, y que tenía un dueño, sin recursos, que se moría de hambre. Al ser un robot kantiano, tenía el imperativo categórico de no robar, y, por tanto, no podía hacer nada, salvo ver morir de inanición a su dueño. Supongamos ahora que este robot está programado con una ética utilitarista. Entonces, el robot robará un pan al panadero para salvar a su dueño.

El robot se dará cuenta de que, si roba el pan, la felicidad de su dueño será mucho mayor que la posible infelicidad del panadero, pues no es lo mismo salvar la vida que perder un pan. De esta manera, el robot robará un pan al panadero todos los días, porque es mayor la felicidad de su dueño que el dolor del panadero. Incluso decidirá robar más panes cada día, para darlos a otras personas que se mueren de hambre, al menos mientras el panadero no se muera también de hambre. El robot verá que la felicidad de todos aquellos que reciben un pan cada día es mucho mayor que el desagrado del panadero. Claramente, salvar las vidas humanas de muchas personas produce en media más felicidad que el daño que se hace a una sola persona, que es el panadero. Posiblemente el panadero no esté muy contento con el utilitarismo.

¡Menudo dilema! Con la ética de Kant matamos de hambre al dueño del robot; con el utilitarismo cabreamos a los panaderos. Resolveremos este dilema en el capítulo siguiente.

De momento, lo relevante con el utilitarismo, a pesar de sus posibles dilemas, es que, conceptualmente hablando, resulta fácil de aplicar en la inteligencia artificial, porque consiste en calcular. En el capítulo 2 hablábamos de la

fórmula del deseo, la llamada «utilidad esperada», que es el producto de la probabilidad por la utilidad. La inteligencia artificial está programada para actuar según aquello que arroje la máxima utilidad, que es una forma de decir que la inteligencia artificial toma la decisión que produce lo que más se desea. Hablamos de utilidad, pero ¿estamos hablando de la misma idea de utilidad que los utilitaristas? Sí y no.

Hablamos de lo mismo en el sentido de que en ambos casos se intenta cuantificar una situación en función de un valor llamado utilidad. No es lo mismo en el sentido de qué mide ese valor de utilidad. En el caso de los utilitaristas ese valor de utilidad es la felicidad. En el caso de la inteligencia artificial, el valor de utilidad es lo que cada diseñador determine: puede ser el número de *likes*, o la cantidad facturada por contenidos. Es la idea de deseo que tenga cada uno. También se diferencia en el cálculo. Para los utilitaristas, la utilidad, que es la felicidad, debe buscar el máximo para la mayoría. Sin embargo, para la inteligencia artificial, la utilidad, que es lo que cada diseñador decida, debe buscar el máximo solo para lo que su diseñador decida. Parece que está bien o mal según lo que cada uno decida. ¡Pues claro, dice Nietzsche!

## Lo que nosotros digamos

### *Nietzsche: no hay moral*

—Pero ¿quiénes somos nosotros para determinar lo que está bien o lo que está mal? —irrumpió Nieves, quien siempre mantenía una visión escéptica de la vida.

Nietzsche es la viva idea de un elefante en una cacharrería. En su caso, en la cacharrería de la moral. Nietzsche irrumpe en la historia de la ética y lo derrumba todo por completo. No deja títere con cabeza, desde los clásicos griegos, hasta Kant o los utilitaristas. Nadie se salva de su lenguaje corrosivo. Porque Nietzsche niega la moral.

No niega que existan ciertas acciones que haya que fomentar y otras que evitar. Lo que discute es cómo se defienden estas acciones. Discute los argumentos que nos llevan a construir una cierta moral, es decir, a definir lo que está bien o lo que está mal. Por eso niega la moral, en el sentido de negar los razonamientos que sustentan toda moral. Propone, en su lugar, una nueva forma de ver la moral[85]. ¿Cómo? Entendiendo la voluntad de poder.

Para Nietzsche, la vida se mueve por la voluntad de poder[86]. Esta voluntad de poder se traduce en apropiación, lesión, conquista de lo extraño y de lo débil, supresión y explotación. Esto es la vida y esto es lo que nos mueve. Esta voluntad de poder, esta necesidad de apropiación y conquista, hace que unos pocos hombres definan lo que está bien o está mal, según su conveniencia, para luego imponerlo sobre el resto de los hombres. Porque la moral es la voz de algunos hombres en el hombre[87]. Hay grupos de personas que dicen que algo es bueno si aumenta su sentimiento de poder y definen como malo lo que les causa debilidad. Estos grupos de personas han sido hasta la fecha de dos tipos: los gobernantes y los esclavos o pobres.

Para los gobernantes, lo que está bien es lo noble, la fuerza o la valentía, porque eso les conviene y aumenta su poder. Por eso, vimos que, para Aristóteles, la magnanimidad, o grandeza de alma, que es reconocerse uno mismo merecedor de grandes cosas, era una virtud; mientras que la humildad estaba cerca del vicio. La moral de la antigua Grecia declaraba como bueno aquello que aumentaba la sensación de poder de los nobles y gobernantes. Es la voluntad de poder de los nobles lo que decide qué está bien. Pero, por otro lado, tenemos la visión del esclavo o del pobre. Este, como ser humano que es, también tiene voluntad de poder. También quiere acrecentar su conquista y dominación. Para ello, ensalza aquello que tiene y que conforma su existencia: la paciencia, la diligencia, la humildad y la amabilidad. Esto es lo que está bien. Ese es su poder. En cambio, la fuerza del gobernante está mal. Es la voluntad de poder de cada grupo lo que define aquello que está bien o está mal y lo impone sobre el resto. Por eso Nietzsche niega la moral, en el sentido de no

aceptar las razones por las que se define el bien o el mal.

Para Nietzsche, debido a la influencia del cristianismo, esta visión del pobre y del esclavo es la que ha triunfado. Por ello, desde entonces, todos los filósofos lo único que han hecho es dar por sentado esta visión, y, movidos también por su particular voluntad de poder, se han dedicado a ofrecer argumentos para defenderla. ¿Kant y su imperativo categórico? Lo único que declara es el orgullo de Kant por saber obedecer y es lo que propone. Es, de nuevo, su voluntad de poder, de imponer lo que para él supone poder, que es obedecer. ¿Los utilitaristas, con su felicidad del mayor número? Lo único que buscan es que la pauta sea aquello que hace feliz a los ingleses. Voluntad de poder anglosajón. Así es la moral con las gafas de Nietzsche.

Y después de este desastre, ¿qué hacemos? ¿cómo construimos una nueva moral?

Aquí viene el problema. Nietzsche entra en la cacharrería de la moral y, después de destrozar toda porcelana ética, vaticina que vendrá una nueva moral, pero que él no puede decirla. ¿Quién es él para hablar de nueva moral? Estaría entonces aplicando su voluntad de poder sobre el resto. En eso hay que reconocer que es consistente. Él no puede hablar de la nueva moral. En su lugar cede la palabra a Zaratrusta, el único autorizado para hablar de un nuevo hombre con una nueva moral[88], un hombre del futuro —no lo busquemos ahora—, un hombre superior: el superhombre. ¡Ay, qué miedo!

Ese nuevo hombre acepta la realidad. Es decir, entiende y admite que la voluntad de poder es la voluntad de la vida. Esa voluntad de poder le llevará a determinar su conducta, sus responsabilidades y sus deberes, pero no con la intención de imponerlos al resto. Porque este nuevo hombre no juzga, no maldice, no regaña. Es lo que llama un hombre espejo, que no se mira a sí mismo, sino que tan solo refleja la realidad. Que dice un «santo sí»[89] a la realidad. Para este nuevo hombre, lo que está mal no es contravenir una norma, sino juzgar, ya sea al otro o a sí mismo.

Si no llegas a entender qué es este nuevo hombre, no te preocupes, es normal. Muchos no lo han entendido y ha sido causa de desgracias. En todo caso, este hombre ya aparecerá. Nosotros somos una especie de hombre mediocre, que solo sabe vivir en un eterno retorno, repitiendo una y otra vez las mismas cosas, como ese ratón que corre por una noria sin avanzar. Por ahora, nos toca seguir como estamos, con las mismas patochadas en TikTok.

### *Pues aplicaremos esta ética mañana*

Hay una cosa en la que coincido con Nietzsche. Dice que todo pensador profundo tiene más miedo de ser comprendido que de ser incomprendido. Ciertamente, pareciera que los filósofos, en general, no quisieran ser entendidos y por ello escriben de tal forma que hay que leer un párrafo tres veces para llegar a entenderlo. Pues esta situación llega a una máxima expresión

en el caso de Nietzsche.

¿Se podría aplicar la ética de Nietzsche a la inteligencia artificial? Imposible. Primero porque Nietzsche no propone ninguna moral, sino que niega y destruye la moral existente. En segundo lugar, porque la solución que propone es para un tiempo futuro, para un hombre nuevo, y difícil de comprender —porque es del futuro—.

Si la *startup* AlegrIA quisiera aplicar las ideas de Nietzsche, no podría emitir ningún juicio sobre ningún contenido. No habría *likes*, ni recomendaciones. Tendría que mostrar un catálogo infinito de contenidos y dejar que cada usuario eligiera el suyo. Pero sin posibilidad de comentar nada. Ningún usuario podría decir este contenido es bueno o malo. ¡Adiós a las redes sociales, quedarían mudas! Tampoco podría juzgar los contenidos que él eligiera. Simplemente serían los contenidos que eligió. Es un «santo sí» a la realidad. Ahora bien, esto no podríamos hacerlo ahora, porque no somos ese hombre del futuro. Es simpático eso de proponer una moral para mañana.

Si no podemos hacer nada con Nietzsche, ¿para qué te lo cuento? Pues porque lo relevante de Nietzsche es el legado que ha dejado. Rompe la moral y dice que la solución vendrá en próximas temporadas de la serie. Vale, pero mientras tanto…, ¿qué hacemos? En ese mientras tanto, lo que ha quedado es ese pensamiento que proclama quién soy yo para decir qué está bien o mal, sin proponer solución. Así lo defiende el personaje de Nieves en esta ficción tan real de la *startup*. Fácil escondrijo para la falta de compromiso.

¿La solución? Tenemos que hablarlo, dice el filósofo Habermas.

### *Ética del discurso: tenemos que hablar*

—Un poco de orden, ¡por favor! —gritó Habana, levantando las palmas de las manos buscando concordia—. Tendremos que buscar una aprobación de la mayoría, un diálogo entre todas las partes, con los usuarios, los productores de contenidos…

Ciertamente, debemos avanzar después de Nietzsche. ¿Pero cómo definir una moral sin caer en la voz de unos pocos hombres en la conciencia de muchos hombres? Pues abandonando la visión de esos pocos hombres e implicando a esos otros muchos hombres. Que no sean unos pocos (los filósofos o los tiranos) los que hablen de normas sobre lo que está bien o mal, sino que lo acuerden las personas afectadas por dichas normas. Surge, entonces, un nuevo problema. ¿Cómo garantizamos que las normas que acuerden esas personas son válidas y no acuerdan cualquier cosa? Será válidas si lo acuerdan dialogando entre todos ellas, pero en un tipo de diálogo especial, llamado discurso práctico. Nace así la ética del discurso (o discursiva) o también llamada ética de la comunicación, desarrollada principalmente por Karl-Otto Apel y Jürgen Habermas[90]. Veamos cómo es ese discurso práctico.

Este discurso práctico tiene que garantizar la idea de universalidad. Esta

universalidad ya la proponía Kant, pero ahora se le da una nueva perspectiva. Vimos que para Kant una norma era moralmente buena si se podía convertir en una ley universal. Sin embargo, esta universalidad es insuficiente, porque parte del pensamiento de una persona, que se pone en el lugar de otros y determina que lo que piensa lo puede aceptar el resto. Estamos, de nuevo, en la voz de unos pocos hombres en la conciencia de muchos hombres. Por ello, Habermas propone esta nueva visión de la universalidad:

> Una norma es universal cuando todos pueden aceptar libremente las consecuencias y efectos colaterales que se producirán previsiblemente de su cumplimiento general para la satisfacción de los intereses de cada uno.

Ahora la norma no es universal porque uno lo piense, sino porque lo determinan todos los afectados, al aceptar libremente las consecuencias de su aplicación. ¿Cómo se consigue esta universalidad? Mediante el llamado discurso práctico. ¿Y cómo es este discurso práctico?

El llamado discurso práctico determina una serie de condiciones para crear una situación de diálogo ideal. Un diálogo en el que el objetivo no sea discutir por discutir, para esto Instagram, ni con el propósito de hacer valer una opinión sobre otra, que para ello está Twitter. El propósito es descubrir en cooperación, entre todos los afectados, lo que es correcto y justo de forma razonada. Para ello, el discurso práctico establece las siguientes condiciones:

- Cualquier persona capaz de hablar y de actuar puede participar en la discusión.
- Cualquiera puede cuestionar cualquier afirmación.
- Cualquiera puede introducir cualquier afirmación en el discurso.
- Cualquiera puede manifestar sus posiciones, deseos y necesidades.
- No se le puede impedir a ningún participante hacer valer sus derechos, establecidos en las reglas anteriores, por medios coactivos originados en el exterior o en el interior del discurso.
- Ningún participante debe contradecirse.
- Si un participante aplica un razonamiento algo para una situación, deberá admitir aplicar el mismo razonamiento para una situación similar.
- Distintos participantes no pueden emplear una misma expresión con significados distintos.
- Cada participante solo puede afirmar aquello en lo que verdaderamente cree.
- Quien introduzca un enunciado o norma que no es objeto de la discusión debe dar una razón de ello.

Esto es un discurso práctico y todo lo demás son conversaciones de café. Solo aquellas normas morales que se establezcan a través de este tipo de discurso práctico se podrán considerar normas válidas. Es lo que Habermas

llama postulado de la ética discursiva:

Únicamente pueden aspirar a la validez aquellas normas que consiguen —o puedan conseguir— la aprobación de todos los participantes en cuanto participantes de un discurso práctico.

Recapitulamos, que ya llevamos muchos principios, condiciones y postulados.

Según la ética del discurso, una norma es válida si cumple el principio de universalidad, pero no el de Kant, sino el que implicaba la aceptación de todos; esta universalidad se consigue si la norma es aprobada en un discurso práctico, el cual, a su vez, se consigue con las condiciones anteriores.

Si yo determino la norma «no debes matar», esta norma no es válida porque no es verdaderamente universal. Es una norma que yo he pensado y, aunque sea de buena fe, he considerado de forma individual que todo el mundo la podría aceptar. Para que sea válida, tengo que reunir a todos los afectados y debatir con ellos esta idea de no matar en un discurso práctico, con las condiciones anteriores. Si dentro de ese diálogo se establece la norma «no debes matar», es entonces cuando se convierte en una norma moralmente válida. Así funciona la ética del discurso.

Vamos a aplicarla a la inteligencia artificial.

### *Tendremos que verlo entre todos*

La ética del discurso no describe una ética particular. No dice debes hacer esto, sino que establece unos pasos, un procedimiento, para determinar si una norma es válida desde un punto de vista moral. Para ello, dicha norma se tiene que haber establecido mediante un proceso de discurso práctico entre todos los afectados por dicha norma. Justo lo que decía Habana, en la *startup* AlegrIA: «Tenemos que llegar a una aprobación de la mayoría». ¡Falso!

La validez de la norma se fundamenta en que sea universal. Esta universalidad es distinta de una aprobación de la mayoría. La norma debe ser aceptada por todos, no aprobada por mayoría; no porque se vote y haya más síes que noes. No es lo mismo universalidad que mayoría. La ética del discurso no habla de contar *likes*. Esta es la trasmutación que se ha producido al llevar la ética del discurso a la esfera pública, en la que se afirma que es válido lo que aprueba la mayoría. Es válido según estatutos, reglamentos o constituciones, pero no necesariamente válido desde un punto de vista moral, desde la moral de la ética del discurso.

Si la *startup* quiere aplicar la ética del discurso para estar segura que propone contenidos éticos, tendría que convocar a todos los afectados por dichos contenidos. Primero los usuarios, pero también a los creadores de contenidos, a los empleados y accionistas de AlegrIA. Una vez reunidos en una sala o foro, se tendría que establecer un discurso práctico en el que todos pudieran hablar y expresar sus razonamientos sin ser coaccionados. Las normas así establecidas

respecto a los contenidos a recomendar serían normas válidas. Pero siempre que se obtuvieran por acuerdo de todos, no por votación. Suena bien, pero parece complicado. ¿Seguro que nadie coaccionaría a otros? ¿Todos hablarían de forma razonada, o se dejarían llevar por sus emociones en un momento dado? ¿Acabaríamos votando para poder terminar algún día?

La ética del discurso tiene una bondad, una dificultad y un riesgo. Su bondad es su visión de universalidad; su dificultad es poder alcanzar ese diálogo ideal que es el discurso práctico, donde uno se expresa sin coacción y de forma razonada; su riesgo, confundir el consenso por la tiranía de la mayoría y llegar a una moral, que ya no será la voz de unos pocos en la conciencia de muchos, sino la voz de la mitad más uno en la conciencia del resto.

## Un cuadro cubista

¿Qué ética podemos aplicar a la inteligencia artificial?

Al final tenía razón el becario de AlegrIA cuando proyectó la pintura cubista *Mujer con abanico* de María Blanchard: esto de la ética es un cuadro cubista. La ética es una realidad compuesta por distintos puntos de vista, cada uno de los cuales tiene sus aciertos y sus errores.

Este capítulo ha sido una degustación de ética para impregnar tu pensamiento de sabor crítico. Te he propuesto un repaso por distintas teorías éticas, sin pretender ser exhaustivo, pues claramente hay muchos más filósofos éticos de los cuales no he hablado. El objetivo era acercarnos a esa cuestión de qué tipo de ética aplicar, si queremos ser éticos con la inteligencia artificial. Porque nos encontraremos con algunas personas y organizaciones que no piensen ni por asomo ser éticos con la inteligencia artificial, pero también con otras muchas que de buena fe sí intenten ser éticos con la inteligencia artificial. En estas últimas surgirá la pregunta de ¿qué ética usamos, de qué modo podemos estar seguros que lo que hagamos con la inteligencia artificial estará bien?

Resulta complicado utilizar una sola visión ética en la inteligencia artificial, porque, ¿cuál elegimos? Podría ser Aristóteles, pero resulta complicado conseguir esa felicidad tan particular llamada *eudaimonía*, basada además en unas virtudes con visión aristocrática. Podemos cambiar de virtudes, y pasarnos a santo Tomás, quien defiende aquellas virtudes que son *lo natural* en nosotros como seres humanos. Pero entonces alguien dirá que lo natural son las emociones y que esto no es más que apetencias, simpatía y utilidad. En la cuestión de la utilidad estarán de acuerdo aquellos que defienden los gustos de la mayoría, como los utilitaristas. Pero estarán en desacuerdo aquellos que, como Kant, defienden partir de unos principios, pero ¿qué principios? Ninguno, para no caer en la tiranía de unos pocos, o bien los consensuados por los afectados siguiendo las pautas de un *diálogo ideal*. Si es ideal, mal andamos para aplicarlo a este mundo tan real. Un jaleo. ¿Qué hacemos entonces?

Este cuadro cubista, lejos de desanimarnos, debe darnos una visión favorable porque habla de nuestra libertad. Toda teoría ética tiene su error y tiene su acierto, su medida y su desmedida, y el reto está en encontrar un equilibrio. Si ser éticos fuera algo matemático, ya habríamos dado con la fórmula de la ética, con la respuesta infalible a toda situación. Pero en ese caso, no tendría mérito ser ético, pues tan solo estaríamos aplicando una fórmula. Resolver un problema aplicando el teorema de Pitágoras no tiene mérito, pues estás aplicando una fórmula que pensó otro. En todo caso, el mérito radica en haber decidido aplicar el teorema de Pitágoras. El mérito radica en decidir, no en calcular. Y decidir es la expresión de nuestra libertad.

Nos enfrentamos a un interesante cuadro ético cubista ante el cual nuestro deber fundamental es ser libres. Decidir en todo momento qué está bien y qué está mal, aún a riesgo de equivocarnos, que lo haremos. Pero aquí coincido con Aristóteles: lo importante es una cierta continuidad de elección durante toda una vida, porque una golondrina no hace verano. Eso es *ser coherente*, que hoy en día se dice *ser auténtico*.

Por ello la razón de ser de este capítulo. Para que no te líen, para que tengas pensamiento crítico y puedas elegir de forma coherente, o auténtica, si lo prefieres. En los debates sobre la ética y la inteligencia artificial aparecerán argumentos como los que hemos visto en la *startup* AlegrIA. Algunos serán bienintencionados, buscando una solución ética, y otros con la intención de embaucar, pues los antiguos sofistas griegos siguen vivos hoy en día, ya sea en su forma más elaborada como demagogos, o en su forma más básica de populismo. Mediante este capítulo espero haberte dado ese espíritu crítico necesario para saber discernir si una propuesta ética en la inteligencia artificial es acertada o no.

Ahora toca saber cómo aplicar ese espíritu crítico. Lo veremos en el siguiente capítulo con la visión de algo que se llama ética aplicada. Veremos qué es la ética aplicada y cómo aplicar la ética aplicada. Parece un juego de palabras, de esos que gustan tanto a los filósofos.

Hablando de filósofos… No me resisto a terminar este capítulo sin hacer un llamamiento. Leyendo a Jürgen Habermas me encontré con este párrafo[91]:

> Las reflexiones propedéuticas que he hecho hasta ahora han servido para defender la posición cognitiva en materia de ética frente a las maniobras diversivas metaéticas de los escépticos axiológicos…

Filósofos del mundo: ¿sería posible hablar de otra manera que todo el mundo pudiera entender a la primera?

**Mis conclusiones. ¿Las tuyas?**

En este capítulo hemos abordado la siguiente cuestión: ¿Qué ética podemos aplicar a la inteligencia artificial? Estas son mis propuestas de respuesta:

- La ética es la rama de la filosofía que se ocupa del hecho moral.
- El hecho moral es una dimensión de la vida humana compartida por todos, que consiste en la necesidad inevitable de tomar decisiones y llevar a cabo acciones de las que tenemos que responder ante nosotros mismos y ante los demás, para lo cual buscamos orientación en valores, principios y preceptos.
- Toda moral busca dar respuesta a dos grandes cuestiones: cómo tener una vida buena y cómo ser justos. Es decir, cómo tener una vida que dé sentido al destino de una persona y cómo buscar un equilibrio entre la vida en común, y nuestra autonomía como individuos.
- Muchas teorías éticas han intentado dar respuesta a estas cuestiones, que podemos utilizar en la inteligencia artificial. En esta tabla tienes un resumen de las teorías éticas que hemos visto y su posible aplicación en el ejemplo de la *startup* AlegrIA que utiliza la inteligencia artificial para recomendar contenidos.
- No hay una única teoría ética que podamos aplicar. Cada una de ellas tiene su acierto y su error.
- Nos toca decidir qué aplicar buscando la coherencia. Eso es lo bueno, porque expresa nuestra libertad.

Estas son mis conclusiones, aplica tu espíritu crítico y saca tu perspectiva en este cuadro cubista.

| | **Su teoría en un "twitt"** | **Aplicación en AlegrIA** | **Aspectos positivos** | **Aspectos negativos** |
|---|---|---|---|---|
| Aristóteles | Conseguir la eudaimonía a través del ejercicio de la virtud | Ofrecer aquellos contenidos que hagan más feliz a un usuario | Basta con conocer los gustos de cada usuario | Habría que ofrecerle esos contenidos con moderación y prudencia |
| Santo Tomás | Conseguir la bienaventuranza, que conocemos por la Ley Natural | Ofrecer aquellos contenidos que cumplan las virtudes teologales | Los criterios son únicos (virtudes teologales) y son lo natural | Puede no existir consenso en que sea lo natural |
| Hume | La moral no se razona, es cuestión de aprobar o desaprobar un hecho | Ofrecer contenidos que gusten más, sin buscar razón alguna | Simplemente ofrecer contenidos con más *likes* | Se podría llegar a contenidos inadecuados, si tienen más *likes* |

| | | | | |
|---|---|---|---|---|
| Kant | Lo correcto es actuar bajo un imperativo categórico | Ofrecer contenidos que cumplan las máximas de universalidad, dignidad y autonomía | Los criterios son fijos (imperativos categóricos) | Hay que determinar esos criterios y su aplicación puede llevar a injusticias |
| Utilitaristas | Está bien lo más útil, que es lo que da la mayor felicidad a la mayoría | Ofrecer los contenidos que más gustan a la mayoría | Ofrecer contenidos con más *likes*, porque eso es lo más util | Tiranía de la mayoría y puede llevar a injusticias |
| Nietzsche | No hay moral. Toda moral es voluntad de poder | Dejar que cada usuario elija su contenido, sea cual sea | No hace falta programar código | Ese usuario debería ser un nuevo hombre del futuro |
| Ética del discurso | Validar las normas en un discurso práctico | Acordar entre todas las partes los contenidos a proponer | Todos los implicados estarían conformes | Resulta complicado llegar a un discurso práctico perfecto |

***Latas de sopa Campbell*, Andy Warhol, 1962. Museo de Arte Moderno de Nueva York.**

# Cómo no ser una sopa de datos

Popular (diseñado para un público masivo), transitorio (solución a corto plazo), prescindible (se olvida fácilmente), de bajo coste, producido en masa, joven (dirigido a la juventud), ingenioso, sexy, efectista, glamuroso, gran negocio.

Así definió el artista inglés Richard Hamilton el arte pop[92] y así es, en parte, la inteligencia artificial: por lo que tiene de ingeniosa, sexy, efectista, glamurosa y, sobre todo, gran negocio.

A comienzos de la década de 1960, Andy Warhol se convirtió en el artista pop más reconocido gracias a su obra *Latas de sopa Campbell*. En realidad, no es un cuadro, sino 32 lienzos que representan los 32 sabores de sopa de la marca Campbell. ¿Por qué pintar 32 latas de sopa? Porque eran algo cotidiano y reconocible, porque no significaban nada, porque las odiaba, o quizás fue algo aleatorio, o, quién sabe, si la mezcla de todo ello.

Warhol quería pintar arte pop, pero de forma única. Por aquel entonces, a comienzos de la década de 1960, otro artista pop, Roy Lichtenstein, pintaba sus famosos cuadros inspirados en tiras cómicas. Warhol quería destacar en el arte pop, hacer algo que realmente impactara y a la vez que fuera totalmente diferente de lo visto hasta entonces. Que nadie pudiera decir que su obra se parecía a la de otro pintor.

Pero no encontraba nada nuevo y Warhol se sumía en la depresión. Fue cuando sucedió una velada en la casa de Warhol con unos amigos y Muriel Latow, dueña de una exquisita galería de arte en el Upper East Side de Manhattan, que lo cambió todo. Warhol estaba desesperado y le insistió a Muriel que le diera una idea para pintar algo distinto. Muriel le dijo que aquello le iba a costar 50$. Warhol accedió, le entregó un cheque por esa cantidad y luego le espetó «¡dame una idea fabulosa!».

A partir de este punto, los relatos del evento divergen ligeramente, según cada asistente en la casa de Warhol[93]. Uno de los presentes, cuenta que Muriel le preguntó a Warhol por aquello que más odiaba. En la respuesta salieron las sopas de Campbell y fue cuando Muriel le preguntó «¿cuál?»,

«¡todas!», respondió Warhol. «¿Por qué no lanzarse, comprar una de cada y pintarlas todas?» fue la sugerencia final de Muriel.

La propia Muriel, sin embargo, en una entrevista indicó que, tras recibir el correspondiente cheque de 50$, ella le dijo a Warhol que «pensara en lo más común cotidiano que fuera reconocible de forma inmediata. Algo como una lata de sopa Campbell; de hecho, ¿por qué no una lata de sopa Campbell?». La podría pintar en varias permutaciones. La única pregunta de Warhol fue «¿cómo de grandes?» «De mi altura», respondió Muriel.

Años más tarde, en una entrevista le preguntaron al propio Warhol si la razón de pintar latas de sopa era porque las comía todos los días cuando era pequeño. Él respondió que sí, que tuvo sopas Campbell cada día, durante 20 años. En otra ocasión, un amigo de Warhol, que se moría de cáncer, le hizo la misma pregunta sobre la causa de pintar latas de sopa y el artista contestó que quería pintar la nada; que estaba buscando algo que fuera la esencia de la nada, y eso era aquello. Por último, en una entrevista de radio, el entrevistador estuvo indagando con Warhol sobre si su obra de sopas Campbell tenía algún significado social, si con ello estaba pintando algún aspecto de nuestra cultura. Warhol contestó lacónicamente que no, y concluyó que las latas fueron un objeto elegido al azar.

Tenemos, pues, una obra maestra sin saber con precisión la razón de su existencia: pintar algo cotidiano, algo que odiaba, algo que es la nada o algo aleatorio. Sea cual fuera la causa, la obra *Latas de sopa Campbell* es la esencia del arte pop.

El arte pop es la exaltación de los objetos comunes para convertirlos en obras de arte. Los artistas pop buscaban eliminar los límites entre un supuesto arte de cultura elevada y un arte de cultura baja. Para ello, tomaban cualquier elemento común de la cultura de masas, ya fueran latas de sopa, comics o anuncios publicitarios y lo elevaban a obra de arte. Era un arte dirigido a ese público masivo, popular, al cual se le podía atraer la atención sobre una cosa novedosa y al día siguiente sobre otra, porque ya había olvidado la primera. Bastaba con mostrar algo con la apariencia de joven, ingenioso, sexy, efectista, glamuroso. Por ello se convirtió en un gran negocio.

Hay una inteligencia artificial que también es un gran negocio. También es joven, ingeniosa, efectista y glamurosa. La inteligencia artificial que es el ingenio que mueve los motores de nuestras aplicaciones móviles para que nos muestren contenidos atractivos, sexys y glamurosos. Y si los olvidamos pronto, mejor, porque así tiene la oportunidad de mostrarnos otros contenidos. Es la inteligencia artificial pensada para el público masivo. Para los millones de ciudadanos que pasan por un punto de control, que puedan caminar por la calle, o que pueblen el mundo ideal del metaverso.

El público masivo es el objeto de esta inteligencia artificial, pero con una diferencia respecto al arte pop. El arte pop tomaba elementos de la cultura de

masas y las elevaba a arte. Esta inteligencia artificial de la que hablo, toma a la masa, a los millones de usuarios de un sistema o una aplicación, y los convierte en datos. La pintura *Latas de sopa Campbell* de esta inteligencia artificial se llama *Latas de sopa de usuario*. Hay una inteligencia artificial para la cual nosotros somos como latas de datos, los cuales se analizan y se clasifican. Es la inteligencia artificial pop.

En el capítulo 1 vimos que muchos peligros que nos anuncian de la inteligencia artificial ya los hemos vivido en el pasado y, de momento, aquí seguimos. Otros riesgos los arreglaremos, como lo hemos hecho antes. Pero lo que tenemos que vigilar son aquellos riesgos que forma parte de los valores que transmite la inteligencia artificial: el valor de la autonomía y el valor de la justicia. Consiste en evitar ese despotismo digital que busca lo mejor para nosotros sin contar con nosotros, y en evitar la injusticia, ya sea por sesgos en esos botes de datos o por decisiones basadas en cálculos matemáticos de datos históricos que cometen un doble error, pues la justicia nunca es un cálculo matemático y siempre podemos cambiar nuestra historia.

¿Cómo lo hacemos? ¿Cómo evitamos estos riesgos? Si en el capítulo anterior hemos visto que no hay una única ética suficiente, que cada una tiene su acierto y su fallo, ¿qué ética aplicamos? Probemos con la ética aplicada. No es un juego de palabras. La ética aplicada es una guía para utilizar de forma práctica distintas teorías éticas. Una guía que podemos utilizar para evitar estos riesgos éticos de inteligencia artificial pop.

En particular, podemos volver nuestro Comité de Dirección de la *startup* AlegrIA y ver cómo resolver su dilema de ser una empresa que utiliza la inteligencia artificial de forma ética para ofrecer contenidos a sus clientes. Vamos a ver si podemos dar una alegría a AlegrIA.

## Ética aplicada

### *Entre principios y consecuencias*

En el capítulo anterior hemos hecho un breve repaso por algunas teorías éticas significativas. Ahora vamos a hacer algo que le encanta a la inteligencia artificial, que es agrupar. Vamos a intentar encontrar algún elemento en común entre todas estas teorías éticas para luego agruparlas. El objetivo es entender las teorías éticas desde una visión superior.

Comencemos por Aristóteles, que fue el primero en nuestro repaso. Recordarás que para el filósofo griego el bien moral es aquello a lo que tiende cada cosa, que es su *fin*. En el caso particular del hombre, su fin es la felicidad. Por tanto, el bien moral es la felicidad del hombre. Algo es bueno si su fin es la felicidad. Es verdad que era una felicidad algo especial, llamada *eudaimonía*, pero felicidad, al fin y al cabo.

Para santo Tomás, el *fin* del hombre es también la felicidad, pero ahora una felicidad suprema, llamada bienaventuranza, que es la contemplación de Dios en su esencia. En este sentido, está bien aquello cuyo *fin* tiende a la contemplación de Dios. Santo Tomás difiere con Aristóteles en cómo se alcanza esa felicidad y por ello aporta nuevas virtudes que nos llevan a esa bienaventuranza, pero comparte con él la idea de un *fin* que mueve nuestros actos y que determina lo que está bien o está mal.

Damos ahora un salto hasta los utilitaristas. Bentham incorpora el término utilidad para hablar de aquello que nos produce beneficio, ventaja, placer, bien o, en definitiva, felicidad —otra vez—. De esta forma, si algo te causa mucha felicidad, entonces tiene mucha utilidad, y más utilidad todavía, si causa la felicidad de la mayoría. Ahora, por tanto, algo está bien si su *fin* es aumentar la utilidad, que es aumentar la felicidad de la mayoría.

En los tres casos he destacado la palabra fin. Estas filosofías éticas se preocupan por el fin que se persigue con una determinada acción, y determinan que esa acción está bien si con ella se tiende a dicho fin. En los tres casos que hemos visto, ese fin es la felicidad, de una forma u otra. Estas éticas se preocupan por las consecuencias de nuestros actos. Si la consecuencia es la felicidad, en la forma que cada ética determine, nuestros actos son buenos. Este tipo de ética que define el bien moral en función de los fines o de las consecuencias obtenidas se denomina ética teleológica —el nombre es rarito, sí —.

Pero Kant no está de acuerdo.

Kant fundamenta su moral en lo que llamamos imperativo categórico. Tenemos que hacer las cosas por un deber incondicionado. No obramos a condición de obtener un fin (por ejemplo, la felicidad), sino porque hay algo que nos obliga desde el comienzo, desde el principio. Esta obligación se

fundamenta en unas leyes o en unos *principios*, que en el caso de Kant son la universalidad, autonomía y dignidad. Una acción está bien si cumple con esos principios. Según Kant, el bien moral no es conseguir un cierto fin u obtener una cierta consecuencia, sino cumplir con una serie de *principios*. Este tipo de ética basado en los principios se denomina ética deontológica —con un nombre no menos raro—.

Si lo vemos como un sano e intelectual combate filosófico, en una esquina del cuadrilátero tenemos las éticas deontológicas, que establecen unos principios que luego se deben seguir a toda costa, y en la otra esquina la liga de las éticas teleológicas, que establecen una consecuencia y deciden que algo está bien si se obtiene dicha consecuencia. ¿Quién ganará en este apasionante combate filosófico?

Según la experiencia de la *startup* AlegrIA que vimos en el capítulo anterior, ninguna de las dos éticas gana. La ética utilitarista es el ejemplo más claro de ética teleológica. Cuando intentábamos aplicarla al caso de un robot que decide robar pan para salvar vidas humanas, vimos que era la desgracia de los panaderos. Pero si aplicábamos Kant, tampoco salían las cuentas. En este caso, el robot veía cómo su dueño se moría de hambre por evitar robar una barra de pan y violar así el principio inscrito de no robar.

Pero vamos a intentar un último asalto, a ver si conseguimos un ganador.

### *Dime a quién atropello*

En 1967, la filósofa Philippa Foot propuso el llamado dilema del tranvía[94]. En él supone que un tranvía se encuentra fuera de control, sin posibilidad de frenar, y se dirige hacia un tramo en el que se encuentran trabajando cinco operarios en la vía. Si nada lo impide, arroyará y matará a estos cinco trabajadores. Antes de dicho tramo se encuentra un cambio de agujas que permite cambiar el trayecto del tranvía a una vía secundaria. Por desgracia, en esta segunda vía se encuentra un trabajador operando también, y si cambia de vía, de igual forma será arroyado.

Por cuestiones del destino —y porque si no, no tiene gracia—, ocurre que tú te encuentras en el cambio de agujas. Si no haces nada, el tranvía continuará su camino y matará a cinco personas; si accionas el cambio de agujas, el tranvía pasará a la vía secundaria y acabará matando a una persona. ¿Qué haces? ¿Accionas el cambio de agujas y muere una persona o no lo accionas y mueren cinco? Pues depende de cómo sea ese filósofo que llevas dentro.

Si tu filósofo interior es utilitarista, claramente accionarás el cambio de agujas. Hemos visto que los utilitaristas dicen algo está bien si su fin es aumentar la utilidad, que es aumentar la felicidad de la mayoría. Siendo triste que muera una persona, es más triste que mueran cinco. Por lo tanto, tiene más utilidad salvar a cinco personas —hay más gente feliz— que salvar a una persona. Como buen utilitarista, aunque no lo sepas, accionarás el cambio de

agujas. Si además eres banal, grabarás un TikTok para mayor gloria y recibirás muchos *likes* de los parientes de los cinco trabajadores salvados —en eso consiste la idea de utilidad de los utilitaristas—.

Pero quizás en tu interior se encuentre un pariente lejano de Kant. En ese caso, no accionarás el cambio de agujas. Para Kant obramos bien si actuamos según un imperativo categórico basado en unos principios. Estos principios deben cumplir tres máximas: universalidad, dignidad y autonomía. Vamos a centrarnos en la máxima de la dignidad. Kant decía: «Obra de tal modo que trates a la humanidad, tanto en tu persona como en la de cualquier otro, siempre como un fin y nunca solamente como un medio». La máxima de la dignidad nos impide usar a cualquier ser humano como un medio para obtener algo, con independencia de que ese algo sea tan loable como salvar cinco vidas humanas. No podemos disponer de la vida del pobre trabajador que opera en la vía secundaria, y decidir por él que debe salvar a cinco compañeros. No podemos usarle como un medio para salvar a otros. «¡Está bien esto de ser el héroe porque lo decida otro!», bien podría decir el trabajador afectado. En consecuencia, como buen kantiano no accionarás el cambio de agujas. Quizás no te hagas un TikTok, porque esto del imperativo categórico, las máximas y la dignidad es muy complicado de explicar, y poco tiene que hacer frente al heroísmo mediático de salvar cinco vidas.

Este dilema del tranvía lo podemos trasladar a un vehículo de conducción autónoma, que se encuentre ante el dilema de tener que atropellar a unos viandantes o a otros. Así lo hicieron un conjunto de instituciones, entre las que se encuentra el Instituto de Tecnología de Massachusetts (MIT), cuando crearon una plataforma web llamada Moral Machine[95]. En esta página web plantean distintas alternativas al dilema del tranvía para el caso de un vehículo autónomo conducido por inteligencia artificial. Dado el supuesto de un vehículo autónomo sin frenos y sin posibilidad de salirse de la calzada, se encamina hacia un paso de cebra en el que se encuentran distintos peatones, ya sea en su carril o en el carril contrario. El dilema radica en atropellar a los transeúntes de uno u otro carril.

El objetivo de esta plataforma, en palabras suyas, es «construir una imagen usando *crowdsourcing* de la opinión de los humanos sobre cómo las máquinas deben tomar decisiones cuando se enfrentan a dilemas morales». Es decir, hacer una encuesta —palabra más rancia que *crowdsourcing*— de a quién cree usted que debemos matar, para luego programar así la inteligencia artificial. Según los últimos resultados publicados en 2020, en torno al 80% de los participantes opinan que lo mejor es matar al pobre trabajador de la vía secundaria[96].

¡Ya tenemos ganador del combate filosófico entre éticas deontológicas, representadas por Kant, y éticas teleológicas, representadas por los utilitaristas! Por aclamación popular, ¡ganan los utilitaristas!

Luego ya no tenemos dudas. No hace falta darle muchas vueltas a esto de

pensar en una ética para la inteligencia artificial. La mayoría de la población opina, sin saberlo, que la mejor ética es la utilitarista. Basta entonces con programar la inteligencia artificial para que haga sus cálculos de utilidad y decida según la mayor felicidad de la mayoría.

¡Un momento! Permítaseme, cuestionar el resultado.

Hay un matiz muy importante entre el dilema del tranvía propuesto por Philippa Foot y el dilema de un vehículo autónomo con inteligencia artificial respecto a quién atropellar. El matiz eres tú. En el dilema del tranvía, tú estás en el cambio de agujas y tú decides, según tus razones, valores o creencias. En el vehículo autónomo, tú estás dentro del coche, pero tú no decides. La solución del dilema va a depender de si el ingeniero o ingeniera diseñadores son deontológicos (por ejemplo, kantianos) o teleológicos (por ejemplo, utilitaristas) y así programarán la inteligencia artificial del vehículo en consecuencia. O bien, si quieren quitarse problemas, programarán lo que opina la mayoría. Pero en cualquiera de estos casos, tú no decides. En la conducción de un vehículo autónomo, tú estás en el cambio de agujas y otro decide por ti. Así puede ocurrir en cualquier sistema operado por inteligencia artificial.

Ya vimos en el capítulo 4 que ser ético consiste en tomar decisiones y responder de esas decisiones ante nosotros mismos y ante los demás. No podemos delegar nuestra autonomía, si lo hacemos, abandonaremos el humanismo para convertirnos en mecanismo.

¿Cómo evitarlo?

### *Por qué combatir*

El primer paso radica en no enfrentar las éticas de los principios (deontológicas) con las éticas de las consecuencias (teleológicas). Durante los párrafos anteriores he planteado una falacia, es decir, he usado un argumento que parecía válido, pero no lo era. He usado la falacia de la falsa disyunción, que consiste en oponer dos términos cuando estos, o bien no son excluyentes (se pueden contemplar ambos), o no son exhaustivos (hay más opciones). A los tiranos les gusta mucho este tipo de falacia, cuando, de una u otra forma, plantean el argumentario de conmigo o contra mí.

He supuesto que las éticas de los principios y las éticas de las consecuencias son excluyentes, cuando no lo son. Aunque parezca mentira, Kant y los utilitaristas pueden llegar a un acuerdo. Así lo expuso Max Weber en su famoso ensayo *La política como vocación*[97], donde habla de la ética de la convicción y la ética de la responsabilidad, que es lo mismo que la ética de principios y la ética de consecuencias.

Lo relevante de Weber fue defender que, al menos en el ámbito político, un dirigente no podía conducirse por una ética de principios, sino por una ética de las consecuencias. Un político es responsable de sus decisiones y, por tanto, debe medir más las consecuencias de sus acciones, que basarse en la radicalidad

de sus convicciones. Esta afirmación parece defender la visión del político que carece de principios y para el cual todo vale, porque depende de lo que se consiga. No es así, que no se alegren los políticos veleta, si se creen refrendados por un filósofo.

La propuesta de Weber, que parece tan exagerada, no muestra estas dos éticas de manera excluyente. Weber plantea la ética de las convicciones y la ética de la responsabilidad como dos opciones que se complementan y que guían a la persona sensata. Lo mejor es escuchar al propio Weber, que lo expresa muy bien:

> […] es muy conmovedora la actitud de cualquier hombre maduro […], que siente con toda su alma la responsabilidad por las consecuencias y actúa conforme a la ética correspondiente y que, llegado el caso, es capaz de decir: «no puedo hacer nada más, aquí me detengo». Siento que esto es algo realmente humano […]. Desde este punto de vista, la ética de la responsabilidad y la ética de la convicción no son términos opuestos entre sí; son elementos complementarios que deben concurrir a la formación del hombre auténtico.

Puedo usar la ética de las consecuencias, pero en un momento dado debo decir, «aquí me detengo», porque atiendo a mis principios. ¡Esa es la diferencia con la visión del todo vale del político veleta! Los principios y las consecuencias se deben moderar mutuamente. Vimos que, si aplicamos solo la ética de los principios, podemos ver cómo alguien se muere de hambre por no robar un pan; y si solo pensamos en la mejor consecuencia para la mayoría, los panaderos se arruinan. Probemos a usar lo bueno de cada ética. La ética de los principios debe poner unos límites intraspasables para obtener un resultado y la ética de las consecuencias pide cierta flexibilidad en la aplicación de los principios[98].

Encontramos una salida a este combate filosófico. Kant y los utilitaristas se dan la mano. Podemos usar ambas éticas: las éticas de los principios y las éticas de las consecuencias. La siguiente cuestión es cómo articular ambas éticas para obtener un equilibrio adecuado. La solución es la ética aplicada.

### *Los filósofos se dan la mano*

La ética aplicada no es una nueva teoría ética. Es un marco que combina distintas éticas para ayudarnos a pensar y poder tomar decisiones en ámbitos prácticos. La ética aplicada se utiliza para encontrar soluciones prácticas a los dilemas éticos que puedan surgir en disciplinas como la medicina, la investigación biomédica, la empresa o los medos de comunicación. De hecho, la ética aplicada nació con el llamado *Informe Belmont*[99] para aplicar este marco a la bioética. ¿Por qué no probar la ética aplicada en la inteligencia artificial?

La ética aplicada parte de la visión de Max Weber de intenta armonizar ese

entendimiento entre las éticas de los principios y las éticas de las consecuencias. Kant y los utilitaristas se pueden dar la mano, pero hace falta un árbitro que medie entre ellos. Las éticas de los principios y las éticas de las consecuencias son los dos extremos y un comportamiento moderado debe saber moverse entre ambos, sin alcanzar ninguno. ¿Cómo? Con la mediación de las virtudes. ¿Cuáles, las de Aristóteles, las de santo Tomás? No necesariamente.

Recuerda que para Aristóteles la virtud de una cosa es la excelencia o el ejercicio correcto de la función de esa cosa. Según Aristóteles, la función del ser humano es la felicidad (*eudaimonía*), pero para santo Tomás, es la bienaventuranza, o contemplación de Dios en su esencia. En ambos casos, la virtud es lo que ayuda a la excelencia en dichas funciones, según la visión de cada uno. Aristóteles define unas virtudes tales como el valor, la templanza, la generosidad, la gentileza o la grandeza de alma, mientras que santo Tomás aporta las virtudes teologales que son la fe, la esperanza y la caridad. Se supone que cada una de estas virtudes ayudan a conseguir esa excelencia en lo que cada uno entiende que es la función del ser humano.

Ahora buscamos una visión más abierta de la virtud, sin pensar en una función particular del ser humano. Nos centramos en la virtud como expresión de la excelencia de la persona. ¿Qué es esta excelencia? ¿qué es una persona excelente o una persona virtuosa? Es aquella que se siente realizada en el sentido de hacer solo aquello que juzga digno de lo que él es o quiere ser. La virtud es un esfuerzo exitoso por ser lo que uno razonadamente desea ser[100]. ¡Ojo! Aquí hay muchas palabras que merecen atención.

Esta excelencia habla de sentirse realizado, expresión que se utiliza mucho hoy en día. Pero no es sentirse realizado para hacer aquello que tenga más corazoncitos en Instagram o hacer aquello que otros esperan de ti. Es un sentirse realizado según la idea razonada de lo que uno es o quiere ser y que solo se consigue con esfuerzo. Razonar y esfuerzo. Dos palabras con pocos *followers* hoy en día. Por ello ser virtuoso es tan complicado y escaso.

Al hablar de la definición de ética decíamos que toda moral busca saber qué es eso llamado la vida buena. Si la virtud es el ejercicio correcto de la función del ser humano, esa función de forma genérica es la vida buena. Término sencillo, pero complejo de determinar. En palabras de Alasdair MacIntyre[101], un filósofo referente en la llamada ética de las virtudes, la vida buena es la vida dedicada a buscar la vida buena, y las virtudes nos capacitan para entender más y mejor cuál es la vida buena. Puede parecer un juego de palabras (vida buena es buscar la vida buena), pero es más profundo, porque esta definición nos lleva a una idea de vida buena como un proceso, como un camino que, quizás, nunca termine, porque es la búsqueda de esa excelencia a la que uno aspira, considerando sus aspiraciones y sus límites.

Esta idea de vida buena y de virtud es muy abierta y poco concreta. Lo sé. Su aplicación exige reflexión y perseverancia. La alternativa es hacer clic en el

botón «seguir» de alguna moral y seguir sus directrices y virtudes. Pero hacerlo de forma seria y coherente también exigirá esfuerzo. Para algo más liviano y directo siempre quedan los libros chusqueros de autoayuda llenos de recetas, tan concretas como ineficaces.

Esta idea de virtud es el componente que demandaba Max Weber para encontrar el equilibrio entre las éticas de los principios y de las consecuencias. Las virtudes son la mediación entre la teoría y la práctica. Sin las virtudes no sirven de nada ni los principios ni buscar buenas consecuencias. Sin esa búsqueda de una vida buena, los principios y las consecuencias solo sirven para escudarse detrás de ellos, porque falta la voluntad sincera de tener verdaderamente en cuenta tales principios y sus consecuencias. Ser ético significa responder a la pregunta ¿por qué hice esto? Si solo atendemos a los principios o las consecuencias, sin la medicación de las virtudes, entonces buscamos una respuesta mecánica amparados en unos códigos de conducta establecidos o en unos resultados estadísticos ampliamente aceptados por todos.

Lo relevante en la ética aplicada, como juego de los principios y las consecuencias, es que en el centro estás tú, como persona que decide, y que media entre esos principios y consecuencias con tus virtudes y tu idea de excelencia. Tú eres el que está en cambio de agujas del dilema del tranvía, y no podemos permitir que esté nadie más.

Tenemos el marco teórico de la ética aplicada. Ahora veamos cómo se aplica la ética aplicada —ahora soy yo el que no podía evitar un juego de palabras—.

### *Aplícate con la ética aplicada*

La ética aplicada es un marco para pensar. Ofrece un método de razonamiento que ayuda a enmarcar un problema concreto y a verlo desde la óptica de la moral, y no desde cualquier otra de sus dimensiones: científica, técnica, social o jurídica[102]. Este marco de reflexión se realiza, como hemos visto, yendo y viniendo entre la ética de los principios y la ética de las consecuencias, con una mediación de las virtudes. Este ir y venir ser realiza según una serie de pasos[103] aplicados a una actividad concreta, por ejemplo, la medicina, la economía o los medios de comunicación. Si aplicamos estos pasos a la inteligencia artificial, quedan de la siguiente forma:

- Determinar claramente el *fin específico* que se persigue con el uso de la inteligencia artificial y por el cual esta cobra *sentido* y *legitimidad social*.

Se debe explicar por qué se está recurriendo a la inteligencia artificial en un cierto sistema para obtener un resultado en una actividad. Qué objetivos o metas se buscan, los cuales deben justificar que se use la inteligencia artificial. Además, estos objetivos deben ser legítimos y aceptables desde un punto de vista social. Este primer punto incide sobre las consecuencias de la inteligencia

artificial. En este caso, sobre las consecuencias intencionadamente buscadas, que, se supone, deben ser beneficiosas de tal forma que se entienda y acepte la opción de usar la inteligencia artificial en ese sistema.

- Indagar qué *principios* debemos incorporar para alcanzar ese fin específico, dentro de una *moral cívica de la sociedad* en la que se inscribe.

Estos fines deben estar guiados y limitados por unos principios. Los fines no se pueden conseguir de cualquier forma y estos principios son los bordes infranqueables que en un momento dado nos permitan decir «aquí me detengo». Como estos principios son los límites, el cómo se determinan es algo relevante. No pueden ser los que a uno un buen día se le ocurran o hechos a la medida de una organización, sino que deben buscar esa universalidad que propone Kant, o más aun la que defiende Habermas que pide implicar a los afectados. Esto puede resultar complicado pues, como vimos en el ejemplo de AlegrIA, significaría incluir en un *discurso práctico* a todos los usuarios de la plataforma de AlegrIA afectados por la inteligencia artificial que les recomienda contenido en un sistema. Para solventarlo, se requieren dos cosas: primero, que estos principios estén acordados entre varias partes, dentro y fuera de la organización, de tal forma que no sean producto de la «lucidez» de un CEO de una compañía que se siente iluminado por la verdad; segundo, y muy importante, que tales principios estén enmarcados dentro de lo que se consideran los valores ampliamente aceptados por la sociedad que recibe dicha inteligencia artificial. Con esto último se busca esa universalidad, al menos en una sociedad específica. No obstante, no está exento de controversia y este tema de los principios lo veremos en detalle posteriormente.

- Conocer la *situación y los medios* por los que se alcanza ese fin específico, mediante *información precisa y clara* de cómo está actuando la inteligencia artificial.

Este es un paso intermedio para el siguiente. El objetivo de la ética aplicada es ayudar a tomar decisiones y para ello es necesario disponer de información precisa y clara. Si utilizamos la inteligencia artificial en un sistema para conseguir unos fines, tenemos que explicar cómo está consiguiendo dichos fines. Esto significa ofrecer pautas a los afectados sobre cómo funciona la inteligencia artificial en ese sistema, cómo está tomando las decisiones para actuar de una u otra forma y qué posibles riesgos pueden existir.

- Dejar la *toma de decisión en manos de los afectados* por la inteligencia artificial, los cuales, con esa información precisa y clara, puedan ponderar las *consecuencias*, sirviéndose de criterios tomados de distintas éticas.

Hasta ahora, con los pasos anteriores, tenemos unos fines, es decir, unas consecuencias buscadas de forma intencionada, y unos principios, que, se supone, son aceptables, al menos, en la sociedad en la que vives. Pero quizás esos principios no te convenzan o los veas con algún matiz particular. También puede ocurrir que, además de las consecuencias buscadas, tú percibas otras consecuencias. Por todo ello es necesario que tú, como usuario de un sistema que utiliza inteligencia artificial, tengas la capacidad de tomar una decisión, al menos en algún grado de libertad. Aquí es donde se indica claramente que tú, como usuario, cliente o ciudadano afectado por la inteligencia artificial, estás en el cambio de agujas y tienes la autonomía de decidir.

Para tomar una decisión acertada, debes contar con información precisa y clara, que es lo que se pide en el paso anterior. Esa decisión la debes tomar considerando esas otras consecuencias que pueda tener el uso en ese sistema de la inteligencia artificial. Volvemos así al punto de partida de las consecuencias, pero ahora con una perspectiva distintas. En el punto uno se habla de los fines o consecuencias que intencionadamente se buscan. Ahora se habla de aquellas consecuencias particulares que a ti te afectan personalmente, posiblemente no buscadas, y que has detectado con el análisis de la información ofrecida. Para determinar estas consecuencias debes tener en cuenta tus criterios morales, que podrás tomar de aquellas éticas con las que más concuerdes, y mediado por las virtudes que hayas decidido tener. Por eso, también, en este libro hemos dedicado cierto espacio a ver algunas teorías éticas aplicadas a la inteligencia artificial, para ir formando ese espíritu crítico moral.

De forma resumida, en versión casi *twitteriana*, los pasos de la ética aplicada en la inteligencia artificial quedan de la siguiente forma:

1. Determinar claramente el *fin específico* que *legitima* el uso de la inteligencia artificial en una actividad.
2. Indagar los *principios* dentro la *moral cívica de la sociedad* en la que se vives.
3. Conocer de forma *clara y precisa* cómo está actuando la inteligencia artificial.
4. Dejar que tomes una *decisión*, ponderando las *consecuencias* que la inteligencia artificial puede tener para ti, considerando *tus criterios éticos*.

De esta manera, vamos y venimos entre los principios y las consecuencias, mediados por las virtudes. ¿Qué virtudes? Las que decida cada uno, con esa visión de excelencia personal y de hacer solo aquello que uno juzga *digno de lo que él es o quiere ser*. No puedo (y no sé si debo), por tanto, darte una lista de las virtudes que hay que seguir. Ése es el esfuerzo personal de cada uno y que te recomiendo hacer, pues las virtudes, eso que somos o aspiramos a ser, son las que nos guían en la toma de nuestras decisiones.

Al menos, sí hay una virtud que recomiendo practicar de forma general y que tomo prestada de la filósofa española Victoria Camps. Me refiero a la

*prudencia*[104], pero con la orientación que le daba Aristóteles y que llamaba *phrónesis*. Ser prudente en el sentido aristotélico consiste en aplicar la razón a la práctica. En saber pensar para saber actuar. Es prudente el que ha aprendido a *razonar adecuadamente*. Este razonar adecuado no lo consigue uno pensando en la soledad de sus aposentos. Además, implica deliberar, ponderar y contrastar opiniones. Porque las respuestas éticas no responden a una ciencia exacta. Esta prudencia, este razonar adecuado, es la virtud que debemos usar en los cuatro pasos de la ética aplicada.

Antes de ver con un ejemplo, cómo utilizar esto de la ética aplicada en la inteligencia artificial, nos tenemos que detener en el paso segundo, que pide establecer los principios de una inteligencia artificial ética. ¿Qué es esto de los principios?

La ética aplicada nació en el ámbito de la investigación médica, con el objetivo de establecer unos límites y no permitir cualquier método de experimentación que implique a seres humanos en favor del avance de la medicina. En 1979 se publicó el llamado *Informe Belmont*[105] que establecía unos principios como fundamento para tales límites. Esto principios son los de respeto a la persona, beneficencia y justicia. El respeto a la persona significa que hay que respetar la autonomía de la persona para que esta decida si quiere participar en un experimento médico y dé lo que se llama su consentimiento informado, una vez que le han explicados en qué consiste el experimento y sus riesgos. La beneficencia exige dos cosas: (1) no hacer daño; y (2) acrecentar al máximo los beneficios y disminuir los daños posibles. El principio de justicia establece criterios para seleccionar a posibles personas para una investigación médica, y evitar, por ejemplo, que se seleccione a grupos sociales desfavorecidos, simplemente porque sean más fáciles de manipular. También establece que el resultado de una investigación médica no debe beneficiar solo a las personas con recurso económicos.

Respeto a la persona, beneficencia y justicia son los principios para la investigación en medicina. Son los límites que nos imponemos para decir «aquí me detengo», llegado el caso, dentro de una investigación médica, aunque esta prometa grandes beneficios para la salud. Pueden parecer grandes palabras, pero tienen su aplicación práctica. ¿Cuáles son los principios para la inteligencia artificial? ¿Qué límites nos imponemos?

## Estos son mis principios, pero tengo otros

### *De principios robóticos a principios para nosotros*

Los primeros principios para la inteligencia artificial que se establecieron fueron las Leyes de la Robótica de Isaac Asimov, que vimos en el capítulo anterior y que por comodidad repito aquí (para que no tengas que buscar en páginas atrás):

- Primera ley: Un robot no hará daño a un ser humano o, por inacción, permitirá que un ser humano sufra daño.
- Segunda ley: Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entrasen en conflicto con la primera ley.
- Tercera ley: Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley.

En esencia dicen que un robot no debe hacer daño, debe cumplir las órdenes y debe protegerse, salvo que esto le haga incumplir las dos primeras obligaciones. Asimov se merece un reconocimiento por ser pionero en pensar en unos principios para la inteligencia artificial. Sin embargo, con permiso de Asimov, adolecen de un fallo: hablan de los deberes para un robot (o para una inteligencia artificial), pero no de los deberes para nosotros, de lo que nosotros debemos hacer con la inteligencia artificial. Se podría pensar que en el fondo sí hablan de nuestros deberes. Si la primera ley dice «un robot no hará daño a un ser humano», en el fondo quiere decir «nosotros debemos programar un robot para que no haga daño a un ser humano». Es verdad, pero es mejor expresarlo de esta segunda forma, porque así se remarca nuestra responsabilidad.

Esta visión de poner el énfasis en nuestra responsabilidad es la que se ha seguido después de Asimov[106], para pasar de unas leyes de la robótica, a unas leyes sobre cómo nosotros debemos diseñar algoritmos para una inteligencia artificial. En 2017 la ACM (Association for Computing Machinery) publicó su declaración sobre Transparencia y Responsabilidad Algorítmica[107]. Su objetivo es regular el diseño y construcción de algoritmos que se utilizan en inteligencia artificial para la toma de decisiones. Intenta evitar que esta toma de decisiones sea opaca, con el propósito de poder determinar si un algoritmo está sesgado o es erróneo en su diseño.

La ACM recomienda que las instituciones deben considerar la toma de decisiones mediante la inteligencia artificial con el mismo rigor y bajos los mismos estándares de calidad que si estas fueran tomadas por personas. Siguiendo el espíritu de Asimov, el objetivo es evitar el daño a las personas,

respetar las leyes y asegurar los derechos y libertades fundamentales, pero ahora poniendo el foco en nuestra responsabilidad en el diseño de los algoritmos. Para ello, la ACM establece los siguientes principios:

- Consciencia. Todos los afectados por un algoritmo inteligente, ya sean los responsables de una organización, los diseñadores, o nosotros como usuarios, debemos *ser conscientes* de sus posibles *sesgos* y de sus potenciales daños asociados.
- Acceso y reparación. La Administración Pública debe fomentar la adopción de mecanismos que permitan *investigar y remediar consecuencias adversas* por decisiones basadas en algoritmos.
- Responsabilidad. Las organizaciones deben *responsabilizarse* de las decisiones tomadas por los algoritmos que utilizan, aunque no sea factible explicar en detalle cómo los algoritmos producen sus resultados.
- Explicación. Se anima a que las instituciones que utilizan la toma de decisiones basadas en algoritmos *den explicaciones* sobre el funcionamiento de sus algoritmos, tanto de los pasos que estos siguen, como de las decisiones concretas que toma. Esto es particularmente importante en el contexto de políticas públicas.
- Procedencia de los datos. Los creadores de los algoritmos deben mantener una *descripción* de la forma en que se recogieron los datos de entrenamiento, acompañada de una exploración de los posibles sesgos inducidos en el proceso de recopilación de datos.
- Auditoría. Los modelos, los algoritmos, los datos y las decisiones deben quedar *registrados* para que puedan ser *auditados* en los casos en los que se sospeche de un daño.
- Validación y pruebas. Las instituciones deben utilizar métodos rigurosos para *validar* sus modelos y documentar dichos métodos y resultados. Se anima a las instituciones a hacer públicos los resultados de dichas pruebas.

Estos principios ya comienzan a hablar de algunos aspectos que suenan bien. Destacan la necesidad de ser conscientes de los sesgos o de los daños adversos, lo cual es empezar a pensar en las consecuencias. Remarcan la responsabilidad final de las instituciones respecto de estas consecuencias. Se pide que se dé información sobre cómo una inteligencia artificial toma sus decisiones, cosa que veíamos en los pasos de la ética aplicada (paso 3). Aunque esto último de una forma liviana. No dice que las instituciones deban dar explicaciones sobre el funcionamiento de un sistema inteligente, sino que animan a que lo hagan. También se anima a que las organizaciones publiquen sus resultados de las pruebas de validación ¿Por qué animan y no obligan? Por dos motivos.

Primero, porque en inteligencia artificial, y en especial si trabajamos con redes neuronales, puede llegar a ser muy complicado saber por qué el sistema

ha tomado una cierta decisión. Esto, sin embargo, no debe evitar el intentarlo, ni olvidar la ausencia de responsabilidad. Por ello, en el apartado de responsabilidad, se indica que las instituciones son responsables, «aunque no sea factible explicar en detalle cómo los algoritmos producen sus resultados». No vale escudarse en que «esto es muy complicado». Efectivamente lo es, pero de igual forma que el desconocimiento de la ley no exime de su cumplimiento, el desconocimiento del resultado de una inteligencia artificial no exime de nuestra responsabilidad.

En segundo lugar, porque es necesario un equilibrio entre la completa transparencia de información y la absoluta opacidad. Además de cuestiones sobre propiedad intelectual, hay un tema de seguridad. Se ha demostrado que, si se revela mucha información sobre cómo ha sido entrenado un sistema inteligente, es posible recuperar los datos de entrenamiento y posteriormente alterar su comportamiento para alegría de los delincuentes[108]. Por otro lado, como ya hemos visto, debemos huir de dar como respuesta "lo dice el sistema", cuando éste ofrece un resultado, sin que el afectado por ese resultado sepa por qué. ¿Cómo podemos encontrar ese equilibrio? Mediante la auditoría por instituciones acreditadas. De esta forma, las organizaciones pueden abrir sus sistemas de forma segura y confidencial y nosotros nos fiamos de lo que dictaminen los señores auditores.

Estos principios, por tanto, no están mal, pero tengo otros.

### *Algoritmos responsables y nosotros también*

Tener principios y no poder verificarlos es como no tener nada. Y para verificarlos, es necesario saber cómo aterrizar esos principios, cómo llevarlos a la práctica.

Decir que tiene que haber responsabilidad en una organización significa que tenemos que nombrar a un responsable en dicha organización y que luego podemos preguntar por él. Pedir que se den explicaciones implica tanto disponer de mecanismos técnicos que nos permitan investigar en las tripas del sistema inteligente que utilizamos como de un cierto procedimiento para dar esas explicaciones de forma que todos las podamos entender. A estos elementos que permiten llevar a la práctica los grandes y hermosos principios es lo que se llaman elementos de control.

Al menos así los llama la organización Fairness, Accountability, and Transparency in Machine Learning (FAT/ML). Desde 2014 viene organizando eventos anuales como un lugar de encuentro de investigadores para encontrar métodos de computación rigurosos que atiendan a los retos éticos de la inteligencia artificial. Son conscientes de que la inteligencia artificial es cada vez más compleja, pero a la vez quieren evitar que lleguemos al recurso irresponsable de decir «el algoritmo lo hizo». Por ello entienden la idea de responsabilidad como la obligación de informar, explicar o justificar la toma de

decisiones mediante un algoritmo, así como de mitigar cualquier impacto social negativo o daño potencial. Con este espíritu tan responsable, ofrecen la siguiente declaración como punto de partida[109]:

> Los algoritmos y los datos que usamos están diseñados y creados por personas: siempre hay un ser humano responsable en última instancia de las decisiones tomadas o informadas por un algoritmo. «El algoritmo lo hizo» no es una excusa aceptable en el caso de que los sistemas algorítmicos cometan errores o tengan consecuencias no deseadas, incluyendo los procesos de *machine-learning*.

El desarrollo de tal premisa se realiza mediante estos cincos principios:

- Responsabilidad: Tiene que existir una *persona con autoridad* para tratar con efectos adversos de un algoritmo sobre las personas o la sociedad.
- Explicación: Explicar de *forma no técnica* la razón de cualquier decisión que tome un sistema.
- Precisión: Identificar *fuentes de error o incertidumbre* de tal forma que se puedan mitigar resultados erróneos.
- Auditable: Permitir que un *tercero* pruebe, entienda y revise el comportamiento de un algoritmo.
- Justo: Asegurar que los algoritmos no favorecen la *discriminación*.

Como los principios sin ejecución es perder el tiempo, estos cinco principios tienen sus correspondientes elementos de control para llevarlos a la práctica. Para que tengas una idea de lo que hablo, en la tabla siguiente te muestro algunos de estos elementos de control asociados a su principio.

| | |
|---|---|
| **Responsabilidad** | |
| Hay una persona con autoridad suficiente para tratar efectos adversos de un algoritmo. | ○ Designar una persona responsable del impacto de un algoritmo.<br>○ Definir un plan en caso de impacto de un algoritmo.<br>○ Desarrollar plan de apagado. |
| **Explicación** | |
| Explicar de forma no técnica la razón de cualquier decisión que tome un sistema. | ○ Definir un plan de explicación.<br>○ Desvelar las fuentes de datos.<br>○ Describir los datos de aprendizaje. |
| **Precisión** | |
| Identificar fuentes de error o incertidumbre para mitigar resultados erróneos. | ○ Evaluar la sensibilidad al error.<br>○ Evaluar la sensibilidad a la incertidumbre.<br>○ Desarrollar un procedimiento de ajuste de datos. |
| **Auditable** | |
| Permitir que un tercero audite el comportamiento de un algoritmo. | ○ Documentar y los algoritmos para poder hacer pruebas.<br>○ Tener datos disponibles para auditoría.<br>○ Permitir auditorías públicas (p. ej. Universidades). |
| **Justo** | |
| Asegurar que los algoritmos no discriminan. | ○ Identificar fuentes de discriminación (sexo, edad, color, nivel de educación, etc.).<br>○ Calcular ratios de error por discriminación. |

Son buenos pasos prácticos en busca de unos principios. Falta por ver qué opinamos aquí, en nuestra amada Europa. Porque nosotros los europeos también tenemos principios.

### *En Europa también hay principios*

La inteligencia artificial debe estar «al servicio de la humanidad y del bien común, con el objetivo de mejorar el bienestar y la libertad de los seres humanos». Así lo refleja la Comisión Europea en sus Directrices Éticas para una Inteligencia Artificial Fiable[110]. Mejores intenciones no se pueden tener. Fueron diseñadas por un grupo independiente de expertos denominado grupo de alto nivel —no es para menos— y posteriormente fueron sometidas a revisión abierta por cualquier organización de la Unión Europea antes de publicar una versión revisada final. Esto es lo que se dice implicar a los afectados, cosa buena y que habría puesto muy contento a Habermas en su visión de situación de diálogo ideal.

Un detalle. La Comisión Europea habla de inteligencia artificial fiable y no tanto de una inteligencia artificial ética. ¿Por qué? Es simplemente que la Comisión Europea entiende por inteligencia artificial fiable aquella que cumple tres requisitos: que sea lícita, ética y robusta. Lícita porque debe cumplir con la normativa vigente; ética porque debe tener valores éticos; y robusta en el sentido de no debe fallar y causar daño. Puede parecer una cuestión de vocabulario, pero si nos fijamos bien, los dos últimos elementos, ética y robusta, atienden a la idea de principios y consecuencias que tanto vengo dando la lata. La Comisión Europea pide que se vigile que la inteligencia artificial tenga unos principios éticos, a la vez que se vean las posibles consecuencias de su actuación. Destellos de ética aplicada.

Con esta perspectiva tan prometedora, la Comisión Europea establece los siguiente cuatro principios éticos de la inteligencia artificial:

- Respeto de la autonomía humana. Si trabajamos con inteligencia artificial, tenemos que mantener nuestra capacidad de decidir y de elegir. La inteligencia artificial no nos puede subordinar, coaccionar, engañar, manipular, condicionar o dirigir —fíjate cuántos verbos— de manera injustificada —y habría que ver cuáles son las causas «justificadas»—. Al contrario, la inteligencia artificial está para potenciar nuestras aptitudes cognitivas, sociales y culturales.
- Prevención del daño. La inteligencia artificial no puede menoscabar nuestra dignidad humana ni nuestra integridad física o mental. Tampoco puede perjudicar el entorno natural. Para ello, debe ser *segura y robusta* desde un punto de vista técnico y no ser diseñada para usos malintencionados.
- Equidad. La inteligencia artificial debe ser *equitativa* en varios sentidos.

Debe garantizar que su coste y beneficio esté bien distribuido. Por ejemplo, no sería equitativo que el coste de un desarrollo de inteligencia artificial lo pagaran unos, pero luego otros se beneficiaran de ello. Debe evitar sesgos injustos o cualquier tipo de discriminación. Y se debe permitir que nos podamos oponer a las decisiones que ofrezca el sistema inteligente.

- Explicación. Las organizaciones que usen inteligencia artificial tienen que poder explicar *cómo funciona* esta y cuál es su *finalidad*.

Obviamente estos principios coinciden, en parte, con los que hemos visto. Primero, porque son razonables. Segundo, porque esto no va de inventar la rueda y ver quién es más original. Esto va de poner orden y por escrito unas reglas de juego que van en sintonía con nuestros valores cívicos.

Al igual que vimos con los algoritmos responsables, si estos principios no van seguidos de elementos prácticos para su aplicación, entonces no sirven para nada. Lo que en algoritmos responsables llamábamos elementos de control, ahora se llaman aquí requisitos clave, y sirven para aterrizar estos principios. Nada menos que siete requisitos clave. Para cada uno de ellos la Comisión Europea propone una serie de acciones concretas de ejecución. Aquí tienes los siete requisitos clave:

1. Acción y supervisión humana. Una decisión no puede estar basada únicamente en procesos automatizados. Por ello, aquí se ofrecen pautas de cómo supervisar su funcionamiento e intervenir si es necesario.
2. Solidez técnica y seguridad. La inteligencia artificial debe estar bien diseñada y no producir errores. Esto es algo obvio, pero no evidente, dada la complejidad actual, y creciente, de la inteligencia artificial. Este punto ofrece pasos para minimizar daños involuntarios y evitar daños inaceptables.
3. Gestión de la privacidad y de los datos. Se debe proteger la intimidad de las personas, lo cual está relacionado con el uso que la inteligencia artificial hace de nuestros datos.
4. Transparencia. Este apartado indica acciones para poder explicar el resultado de un sistema inteligente.
5. Diversidad, no discriminación y equidad. Acciones para evitar el sesgo y garantizar que todos podamos acceder a una inteligencia artificial.
6. Bienestar social y ambiental. Desde un punto de vista ambiental, cómo garantizar que la inteligencia artificial es sostenible. Desde un punto de vista social, cómo vigilar el impacto de la inteligencia artificial en las instituciones públicas y en nuestra democracia. Esto es relevante si aplicamos la inteligencia artificial, por ejemplo, a la distribución de ayudas sociales, o en unas elecciones. Todo un mundo de posibilidades para los tiranos que debemos evitar.

7. Rendición de cuentas. De nuevo, tiene que haber una persona responsable de una inteligencia artificial particular. Además, se establecen mecanismos para que esta se pueda auditar. Y si se produce un daño por una inteligencia artificial, no se puede aducir que «lo dijo el sistema» y lavarse las manos. Hay que compensar por el efecto adverso causado.

Todo esto suena muy bien. Son principios a los cuales nadie se puede negar…, pero que tampoco nadie tiene porqué cumplir. Estos principios de la Comisión Europea para una inteligencia artificial fiable no son de obligado cumplimiento, sino que son directrices, es decir, amables recomendaciones. ¡Oh, vaya! Pero como también lo son los principios que hemos visto en apartados anteriores. Tampoco obligan —un ¡oooooh, vaaaaaya! más grande—.

Es lógico, porque estamos hablando de la ética aplicada, como marco ético de actuación. Si queremos algo más coercitivo, nos tenemos que salir del ámbito de la ética y entrar en el ámbito de la ley. Ámbitos que idealmente deben tener puntos de encuentro, pero que no siempre es así.

Hablando de temas legales, aprovecho para introducir brevemente el marco regulatorio que se vislumbra en el horizonte europeo, y al que luego haré referencia.

### *Incluso habrá regulación europea*

En abril de 2021 la Comisión Europea publicó una propuesta de reglamento como marco jurídico para la inteligencia artificial. Es una propuesta y todavía le queda camino por recorrer. Habrá que ver en qué queda. De momento, una cuestión relevante es que clasifica los sistemas inteligentes en distintos niveles de riesgo, para los cuales aplicarán distintos niveles de obligaciones.

Habrá una inteligencia artificial tan peligrosa, tan peligrosa, tan peligrosa, que estará prohibida. No se podrá utilizar la inteligencia artificial que emplee técnicas subliminales que puedan alterar de forma significativa tu comportamiento (si lo alteran solo un poquito, sin causar daño, estará permitida). Tampoco se puede usar la inteligencia artificial para evaluarte según tu comportamiento social, y que luego eso tenga alguna repercusión en un entorno ajeno. Por ejemplo, si la inteligencia artificial descubre que tuviste una época oscura en la que rompías farolas, no por ello necesariamente has de perder el acceso a una posible subvención. Cada cosa tiene sus derechos y obligaciones y la inteligencia artificial no se puede utilizar para darte o quitarte puntos genéricos por tu comportamiento. También propone que esté prohibida la inteligencia artificial para vigilancia masiva, a menos que exista un motivo manifiesto de riesgo de seguridad. Nada de poner por las calles cámaras con reconocimiento facial sin razón aparente, salvo que —siempre hay un hueco legal— lo apruebe una autoridad judicial, o exista un riesgo inminente, en

cuyo caso el reconocimiento facial masivo se podría implantar de forma inmediata en espera de que finalmente lo apruebe un juez —o no, y habrá que quitarlo, pero el escrutinio facial ya estará hecho—.

Después habrá sistemas de inteligencia artificial de alto riesgo. Es el caso del reconocimiento facial, pero no de forma masiva, que está prohibido, sino en un entorno acotado. También sería una inteligencia artificial de alto riesgo aquella que se utilice para los siguientes ámbitos: la gestión de infraestructuras esenciales, como abastecimiento de agua, luz o calefacción; en sistemas de acceso a la educación, o para evaluar a los estudiantes; en temas de empleo, para contratar o promocionar empleados; en el acceso a servicios públicos, en la gestión de fronteras, en la aplicación de la ley o en procesos democráticos. Como ves, todos temas de mucho miedo si se utiliza la inteligencia artificial sin conocimiento o con fines perversos. El riesgo de esta inteligencia artificial es que caiga en manos de un necio o de un tirano. Estos casos están permitidos, pero tendrían unas obligaciones especiales para garantizar su uso ético. Esperemos que sea así.

El siguiente nivel es una inteligencia artificial de riesgo limitado. Se refiere a aquella que se pueda usar con estos fines: para interactuar con nosotros, como, por ejemplo, los *chatbots*; para detectar de una u otra forma nuestras emociones, o sirva para caracterizarnos socialmente (como lo que vimos al hablar del despotismo digital en el capítulo 1); o aquella utilizada para manipular contenido, como, por ejemplo, para crear videos falsos que parezcan reales (*fake news*). En todos estos casos, tendríamos que estar informados de que se está utilizando o se ha utilizado una inteligencia artificial, para que podamos obrar en consecuencia. Hoy por hoy no es así, y este tipo de inteligencia artificial se esconde a nuestros ojos.

Por último, estaría el resto de la inteligencia artificial, la cual se supone que, al no estar contemplada en el resto de casos, se considera de riesgo mínimo. Corresponde, por ejemplo, a sistemas de filtros de correos *antispam*, o a la generación de imágenes para videojuegos. Para estos casos de momento no se prevén obligaciones. Tan solo la recomendación de mostrar buena voluntad de corazón y de adherirse a principios éticos. ¿Qué principios éticos? Los que hemos visto en este apartado.

Volvemos así a nuestro discurso de los principios éticos. Hemos visto principios para una inteligencia artificial ética de distintas fuentes. En la tabla siguiente te hago un resumen de todos ellos.

**Association for Computing Machinery**

- Conocimiento.
- Auditoría.
- Responsabilidad.
- Explicación.
- Procedencia de los datos.
- Reparación.
- Validación y pruebas.

**Fairness, Accountability, and**

**Transparency in Machine Learning**

- Responsabilidad.
- Justicia.
- Explicación.
- Precisión.

**Comisión Europea**

| Principios éticos… | …que se consiguen mediante |
|---|---|
| Respeto de la autonomía humana. | Supervisión humana. |
| Prevención del daño. | Solidez técnica y seguridad. |
| Equidad. | Gestión de la privacidad y de los datos. |
| Explicación. | Transparencia. |
| Diversidad, no discriminación y equidad. | |
| Bienestar social y ambiental. | |
| Rendición de cuentas. | |

Por fortuna, existe coincidencia en los principios, lo cual transmite algo de confianza, en el sentido de que existe cierta unanimidad. En el caso de la Comisión Europea se distingue entre los propios principios y la forma de llegar a ellos, lo cual aclara los conceptos, que en los otros casos aparecen mezclados. De una u otra manera, son buenas palabras con buenas intenciones.

Debemos trabajar con principios, porque son los que nos guían y ponen límites a las consecuencias, pero hay que reconocer que este tema de los principios éticos plantea algunos problemas: hablan de conceptos abstractos y no dejan de ser palabras de buena voluntad que no obligan. ¿Qué hacemos con esto? Vamos a ver cómo abordar, que no resolver, estos temas.

### *Buenas y grandes palabras*

Las organizaciones están llenas de principios que contienen grandes palabras relumbrantes, tales como justicia, dignidad o respeto. Palabras que luego reproducen en gran formato por las paredes de sus edificios, publican en sus páginas web junto a fotos de inspiración zen, o anuncian por televisión mediante locutores de voces cálidas. Valoro su buena intención. Sin embargo, son buenas intenciones que no siempre obligan y no siempre generan responsabilidad. ¿Cómo obligar a una organización a que sea justa? Ella podrá aducir que ya cumple con la ley. Además, ¿qué es la justicia? ¿Cómo quejarse porque uno siente que su dignidad se ve afectada? Salvo casos fragantes de trato vejatorio, ¿qué es la dignidad humana? Si el uso de un producto te ha causado un daño, ¡ten cuidado! Puede que la organización dueña del producto se olvide por un momento de sus grandes principios y pase a buscar al culpable del daño. Culpable que, bien puedes acabar siendo tú, porque aceptaste los términos y condiciones, que en su artículo 54, en el apartado «Otros casos», se refiere la situación que te sucedió, en la que se indica claramente que es imprudencia de uso y no corresponde compensación; o bien nadie es responsable, porque hay tantos actores implicados que al final las responsabilidades se diluyen y resulta

que la causa está en que la abuela fuma.

También puede darse la circunstancia de que la organización sí quiera resolver la cuestión apelando a sus principios, porque realmente cree en ellos. Conozco casos así. Pero quizás no sepa de buena fe cómo tratar con los conceptos abstractos de los que habla los principios. Para estos casos van los siguientes párrafos.

Uno de los problemas de la ética es que trata con conceptos abstractos. No está claro, por ejemplo, qué es la justicia, qué es la dignidad humana o qué es la verdad. Porque son ideas que en realidad no existen. Por eso son conceptos éticos, porque representan un ideal al que aspiramos. Cuando vemos nuestro mundo hay algo que nos desagrada porque no es como entendemos que debería ser. Quizás lo llamemos injusticia, porque tenemos una idea de una situación distinta en la que debería haber algo que nosotros entendemos por justo. Pero esa idea de justicia no existe, no tenemos referente y en consecuencia cada uno tiene su idea de eso que llamamos justicia. Así nace el conflicto, así nace la ética.

Es fácil ver una mesa y pedir otra igual. Y si crees que la mesa que te han dado no es igual a la que viste, el conflicto se resuelve fácilmente. Se mide y se pesa la mesa, se analiza su material, su color y se puede determinar con precisión si las dos mesas son iguales. No ocurre lo mismo con las palabras abstractas de la ética. No tenemos un referente de pesos y medidas de lo que es un kilo de justicia o un metro de dignidad. Por eso hay debates éticos.

La cuestión se complica cuando además se produce un compromiso entre dos principios éticos. Pensemos, por ejemplo, en un vehículo de conducción autónoma con inteligencia artificial que, por evitar un accidente, elimina la posibilidad de que el conductor pueda maniobrar con el coche y en su lugar maniobra la inteligencia artificial. Para evitar un daño, hemos anulado la autonomía humana. Hemos puesto en compromiso dos principios éticos de la inteligencia artificial: autonomía y prevención del daño. Si ya resulta complicado discutir sobre qué es la autonomía humana —¿realmente somos autónomos, somos libres?— o qué es prevenir el daño —¿qué ocurre si lastimas a una persona para salvarla de un peligro inminente?—, ahora encima tenemos que contraponer estos conceptos y ver quien «gana». Dado este caso, ¿qué es más importante, la autonomía o la prevención del daño? Tenemos un dilema ético. Debate asegurado, manjar de tertulianos.

La solución a estos dilemas nunca está cerrada. De lo contrario, no serían dilemas éticos, serían problemas técnicos con procedimientos de resolución. Vamos a ver tres pautas para, al menos, tener una cierta forma de cómo abordar este compromiso entre principios.

La primera pauta es no crear una jerarquía de principios, de tal forma que unos principios sean más importantes que otros. Esto nos llevaría renunciar de unos principios en favor de otros. Una vez que hemos acordado unos

principios, debemos entender que todos son valiosos, de lo contrario, no los habríamos considerado. Si consideramos, entonces, que todos los principios son valiosos, todos deben ser atendidos. Optar por uno de ellos en detrimento total de otro es siempre una tragedia, porque perdemos irremisiblemente un valor[111]. Siguiendo con el caso del conflicto entre la autonomía y la prevención del daño del ejemplo anterior, esta pauta nos tiene que llevar a evitar argumentos tales como «es más importante la autonomía que la prevención del daño», o al revés. Si se llega a ese punto, no debemos continuar por ahí para no llegar a una jerarquía de valores. Tan solo se podría admitir que debemos renunciar a un principio si se demostrara con toda claridad que no hay otra opción. Pero antes de llegar a este extremo, tenemos que recordar lo que vimos de la falacia de la falsa disyunción. Tendemos a planteamientos binarios, porque son más fáciles de evaluar, pero la realidad es más rica, y siempre debemos explorar si hay una tercera vía que nos evite anular un principio en favor de otro.

La segunda pauta es no debatir sobre los principios de forma teórica. Ya hemos visto que los conceptos éticos son abstractos y cada uno tenemos nuestra visión de cómo deberían ser. Cada uno piensa cómo debería ser la autonomía humana o qué es el daño. Para resolver un dilema concreto real, sacaremos poco en claro si nos dedicamos a debatir sobre la autonomía humana o el daño como conceptos abstractos. Si nos ponemos esquistos como teóricos, podemos llegar a este tipo de debate:

—El daño evitado no habría sido tan grande. El vehículo iba muy despacio.

—Depende, hay daños que parecen pequeños, pero dejan secuelas.

—Pero si el daño es pequeño, la secuela es pequeña.

—Existen muchos tipos de daños, y la secuela puede ser sicológica, que es otro tipo de daño. Además, el daño puede ser subjetivo.

En el vuelo de un suspiro hemos pasado de discutir sobre el daño real evitado a discutir sobre los distintos tipos de daño. Poco importa los distintos tipos de daños que existan en el mundo. Lo importante es evaluar el daño que se evitó y sus consecuencias. Esta es la clave, las consecuencias. Si nos encontramos en un dilema ético, para no debatir sobre teoría, sugiero atender a las consecuencias reales o potenciales. De nuevo nos movemos en ese círculo entre los principios y las consecuencias, mediado por nuestras virtudes.

Por último, la tercera pauta es resolver el dilema mediante un proceso de reflexión razonado, basado en los hechos. La decisión final debe estar suficientemente bien razonada; no podemos apelar al «sentido común» o decir que «es evidente». Si una conclusión es de sentido común o es evidente, costará menos demostrarla. Retemos a demostrarlo al que así opine. Este razonamiento se debe basar en hechos demostrables, no en intuiciones, creencias o suposiciones.

Queda una segunda cuestión a resolver cuando hablamos de estos principios

de la inteligencia artificial: son principios que no obligan y esto puede llevar a eludir toda responsabilidad. La cuestión va más allá de que los principios obliguen o no. Hemos llegado a lo que se denomina responsabilidad sin sujeto[112].

La responsabilidad se asienta en el principio de causa-efecto: es responsable aquel que causó el daño. La idea es muy simple. Dada una situación, empezamos a tirar del hilo hacia atrás, viendo las sucesivas causas con sus efectos, hasta llegar a una causa última. El autor de esa causa es el responsable. Pero en ocasiones la responsabilidad carece de responsable. No es posible encontrar el sujeto último que es responsable de un acto. El mundo en el que vivimos es tan complejo que la cadena de causas y efectos ya no siguen un único hilo. Llega un momento en que un efecto es generado por varias causas, por varios participantes, varios grupos o por la sociedad en general.

Supongamos que una organización relevante (R) utiliza una inteligencia artificial para seleccionar candidatos a puestos de dirección. Esta organización declara que uno de sus principios es la equidad, que se traduce en la falta de sesgos de sus resultados. Pero se descubre que su inteligencia artificial ha favorecido la contratación de hombres para puestos de dirección, en lugar de mujeres, en igualdad de condiciones de conocimiento y experiencia. Se levanta la polémica al grito de: ¡la organización (R) es responsable!

La organización (R) aduce que la causa está en que la base de datos de entrenamiento de la inteligencia artificial tenía sesgos. Pero ellos no hicieron el entrenamiento, lo realizó una prestigiosa consultora de servicios (C). Ahora se grita: ¡la consultora (C) es responsable!

La consultora (C) se defiende diciendo que ellos recibieron los datos de una institución pública (P). ¡La institución (P) es responsable!

La institución pública (P) declara que sus datos se basan en series históricas desde 1950 y que, si los datos están sesgados en favor de los hombres, es porque en el pasado existían más hombres directivos que mujeres. La causa última radica en nuestro pasado de sociedad machista (S). ¡La sociedad (S) es la responsable!

Esta es la responsabilidad sin sujeto. La organización relevante (R) no es responsable, la consultora de servicios (C) tampoco, la institución pública (P), menos. La conclusión innegable es que la sociedad (S) es la responsable. ¡Todos somos responsables!, que es igual a nadie es responsable.

Poco se puede hacer con aquellas organizaciones que busquen eludir su responsabilidad amparándose en este hecho de responsabilidad sin sujeto. Pero hay una posibilidad para aquellas que de buena fe sí se sientan responsables de sus principios. Para ellas, de nuevo, va esta propuesta.

La solución es un cambio de mentalidad. Abandonar la mentalidad de causa-efecto para pasar a una visión de responsabilidad sin culpa. Sentirte responsable, aunque no seas el culpable, el causante último. Para ello, debemos olvidar

plantear la pregunta ¿quién lo hizo?, para responder a la pregunta ¿qué debemos hacer para evitarlo?[113]. Aunque la organización relevante (R) no sea la culpable del sesgo en su inteligencia artificial, si realmente cree en el principio de equidad y se siente responsable moral de lo sucedido, entonces deberá poner todos sus esfuerzos en evitar de nuevo esa situación, al margen de su responsabilidad legal. Lo mismo tendrían que hacer la consultora (C) o la institución pública (P), por ser participantes del hecho. También incluso nosotros, como sociedad (S), aunque resulte más complicado. Ninguno está obligado, porque ninguno tiene culpa, pero nos podemos responsabilizar si de verdad creemos en los principios que proclamamos.

Conceptos abstractos y buena voluntad que no obliga. Estas son las deudas de tratar con principios. Lo que aquí he propuesto, como dije, no lo resuelve, pero permite empezar a abordarlo. No salda la deuda, pero nos da crédito.

Una vez que sabemos de los principios para una inteligencia artificial ética y de cómo lidiar con ellos, corresponde ponerse manos a la obra. ¿Cómo se hace con esto de la ética aplicada en la inteligencia artificial? Para esta suerte de pregunta, veremos una suerte de respuesta.

## Seamos éticos, que no cuesta tanto

### *Una tragicomedia ética*

Recordemos que esto de la ética aplicada, más que una teoría ética en sí, es un marco de reflexión que ayuda a enmarcar un problema concreto para verlo desde la óptica de la moral. Este marco de reflexión consiste en una serie de pasos, que de forma resumida quedan de la siguiente forma:

1. Determinar claramente el *fin específico* que *legitima* el uso de la inteligencia artificial en una actividad.
2. Indagar los *principios* dentro la *moral cívica de la sociedad* en la que se vives.
3. Conocer de forma *clara y precisa* cómo está actuando la inteligencia artificial.
4. Dejar que tomes una *decisión*, ponderando las *consecuencias* que la inteligencia artificial puede tener para ti, considerando *tus criterios éticos*.

Para ver cómo utilizar la ética aplicada en la inteligencia artificial, ¡qué mejor que volver a nuestra *startup* AlegrIA! Recuerda que su Comité de Dirección estaba preocupado por las implicaciones éticas de su algoritmo, basado en inteligencia artificial, que servía para recomendar contenidos a sus clientes. Tuvieron una reunión para decidir qué criterio ético utilizar en el algoritmo y aquello se convirtió en un debate de múltiples ideas éticas, sin solución concreta. Lo único que salió en claro fue el despido del intrépido becario, quien se atrevió a proyectar la imagen de un cuadro cubista para remarcar ese arrebato de visiones éticas. Aquello se convirtió en una tragicomedia: comedia por la locura de propuestas éticas, y drama para el becario despedido. Después de tanto desconcierto con las teorías éticas, podemos ir en su auxilio con esto de la ética aplicada.

La ética afecta a las personas. En esta tragicomedia de AlegrIA tenemos que detenernos a averiguar quiénes son las personas que juegan en esta obra. Ellos serán los implicados o afectados respecto a las decisiones éticas que se tomen.

Tenemos dos protagonistas principales. Por un lado, el propio Comité de Dirección, cuyos miembros de buena fe se interesan por ofrecer una inteligencia artificial ética a sus clientes. Después están los clientes de AlegrIA, que reciben las recomendaciones de contenidos por medio de la inteligencia artificial. Pero puede que haya más actores en el escenario.

La inteligencia artificial es tan compleja de programar que no siempre las organizaciones tienen la capacidad de hacerlo por ellas mismas. O bien prefieren partir de sistemas de inteligencia artificial ya elaborados y adaptarlos a sus necesidades. Por ejemplo, si se quiere utilizar un *chatbot*, no suele merecer la pena programarlo desde cero, y es mejor recurrir a empresas tecnológicas que

ofrecen esa tecnología. Lo único que hay que hacer es adaptar unos de sus *chatbots* ya desarrollados al contenido específico que requiere la organización. Estas empresas tecnológicas también son actores de esta tragicomedia, sin olvidar a los técnicos de la propia organización, que tienen que hacer que todo funcione en sus sistemas.

En el caso de AlegrIA, por complicar la trama e incluir más variables éticas, vamos a incluir a una de tales empresas tecnológicas. Supongamos que su algoritmo de recomendación de contenidos se ha unido a un *chatbot* de una empresa externa tecnológica, el cual tiene la capacidad de analizar las emociones del usuario que solicita un contenido. Este *chatbot* no lo desarrolló AlegrIA, sino que lo contrató a la reconocida empresa de chatbots Dígamebot.

Habitualmente se incorpora a un último actor que es la sociedad en general. Es verdad que la sociedad, sean o no sus miembros clientes de una cierta empresa, se ve afectada por las consecuencias éticas de las decisiones que esta tome. Se ve muy claramente en el ámbito medioambiental. Si una organización contamina, no lo hace solo para sus clientes, sino que afecta a toda la sociedad. De forma similar ocurre con la tecnología, y en particular con la inteligencia artificial. Si disponemos de vehículos autónomos con inteligencia artificial circulando por las calles, toda la sociedad se ve afectada. La sociedad en general suele ser otro actor.

En esta tragicomedia de AlegrIA, el impacto en la sociedad de su algoritmo inteligente junto con el *chatbot* para recomendar contenidos es más reducido. No obstante, lo consideraremos. Los contenidos que visualice una parte de la sociedad, motivados por la inteligencia artificial, pueden afectar al resto de la misma. Máxime si, como cabe esperar, AlegrIA se convierte en una gran plataforma de contenidos con muchos clientes. Por imaginar una locura, supón que todos sus clientes se pusieran a ver vídeos de patochadas descerebradas, motivados por la inteligencia artificial. ¡No quiero ni pensar a lo que llegaríamos como sociedad! Visto así, se presenta:

«SEAMOS ÉTICOS, QUE NO CUESTA TANTO»
Tragicomedia ética en tres actos y un corolario.
La *startup* AlegrIA, que ofrece contenidos multimedia, en su afán por ofrecer un mejor servicio, ha incorporado un algoritmo de inteligencia artificial a su *chatbot* de comunicación para recomendar contenidos más apropiados a sus clientes. El Comité de Dirección de AlegrIA siempre ha manifestado una gran responsabilidad ética en sus servicios que ofrece. ¿Podrá seguir manteniendo su visión ética con la incorporación de la inteligencia artificial?
Elenco de actores
Protagonista 1: Comité de Dirección de AlegrIA.
Protagonista 2: Usuarios de AlegrIA.
Actores de reparto: Técnicos de AlegrIA.
Actriz revelación: Dígamebot, fabricante del chatbot.
Actor secundario: La sociedad global en la que se desenvuelve AlegrIA.

### *Acto I: Una inteligencia artificial legítima*

Determina claramente el fin específico que se persigue con el uso de la inteligencia artificial y por el cual esta cobra sentido y legitimidad social.

En el Comité de Dirección de AlegrIA surgió la pregunta clave. ¿Por qué utilizamos la inteligencia artificial en nuestro sistema de recomendación de contenidos? ¿Por qué usamos un chatbot que puede identificar emociones? ¿Qué fin perseguimos con ello?

No valía cualquier tipo de respuesta a esas preguntas. Si AlegrIA quería mantener su responsabilidad ética, las respuestas debían justificar el uso de la inteligencia artificial, frente a otras tecnologías menos invasivas y estas respuestas debían, además, tener una legitimidad social. Tenían que ser socialmente aceptables.

Sabían que otras empresas del sector utilizaban la inteligencia artificial y el análisis de emociones para mantener a sus clientes adictos a sus contenidos y así facturar más. Conseguir más ingresos era una respuesta que justificaba el uso de la inteligencia artificial. Estaba demostrado que la inteligencia artificial es muy buena clasificando a los usuarios e identificando con precisión sus gustos y preferencias. Así no se podía fallar en la recomendación de contenidos, y el cliente quedaba pegado a la pantalla. Pero aquella respuesta no era legítima desde un punto de vista social. Si así lo fuera, ¿se atreverían estas empresas a publicar claramente esta finalidad?

El Comité de Dirección lo tenía claro. Ser ético significa responder de tus actos, ante uno mismo y ante los demás. En este caso, responder ante los demás era explicar públicamente a sus clientes, y a la sociedad en general, las razones legítimas que los llevaban a utilizar la inteligencia artificial en su algoritmo. ¿Por qué usaban la inteligencia artificial?

Ariadna, la directora ejecutiva de AlegrIA, recordó haber visto una respuesta a esa pregunta en cierta ocasión. Revolviendo por el buscador de Google, dio con ello. Era una aplicación para móvil, creada en la Universidad de Cambridge, para detectar de forma precoz la infección de COVID-19 mediante sonidos de tos, de voz o de respiración. Si te descargabas la aplicación, bastaba con toser frente a tu móvil, respirar fuerte o decir unas palabras predefinidas, y luego indicar si habías pasado la COVID-19. Con esa información, se podría seguir investigando para encontrar signos de COVID-19 de forma temprana. Lo bueno era que en la página web de la aplicación se decía claramente su objetivo: «recoger datos para la detección precoz de la COVID-19 a través de algoritmos de aprendizaje automático, basados en sonidos de voz, de respiración, y de tos»[114]. Estaba justificado el uso de inteligencia artificial y era socialmente legítimo. Ellos, en AlegrIA, tenían que buscar algo similar.

Bastaba con recurrir a aquello que inspiró la creación de AlegrIA: ofrecer

contenidos únicos para gente única; aquel era su lema. Contenidos únicos porque se preocupaban por contenidos, más allá de los puramente comerciales; la gente única significaba el respeto por la individualidad. La inteligencia artificial era la aliada perfecta para esta visión. Mediante el aprendizaje automático se podían conseguir dos cosas: primero, conocer las preferencias de cada usuario, según sus gustos, su experiencia, pero también según sus emociones de ese momento, lo cual atendía más a esa idea de personas únicas; segundo, disponer de contenidos de calidad, analizados según varios criterios, y no solo aquellos con más *likes*. El *chatbot* permitiría que la interacción fuera fácil y rápida. Por último, había que darle al usuario la libertad de poder usar este sistema o no.

Tras horas de debate, pensando bien las palabras, el Comité de Dirección de AlegrIA obtuvo su fin específico y legítimo del uso de la inteligencia artificial para la recomendación de contenidos. Así lo anunció en su página web:

> Usamos la inteligencia artificial para que tengas la posibilidad de disponer de buenos contenidos, avalados por distintos criterios de calidad, conociendo tus preferencias y emociones a través de un chatbot, para que te sea fácil.

Quizás se podría mejorar, pero servía para justificar el uso de la inteligencia artificial. ¿Y la legitimidad social? En AlegrIA veían dos razones: porque la inteligencia artificial permitía combinar criterios de varias fuentes, junto con el conocimiento del usuario, sin caer en la tiranía de la mayoría, o en el adoctrinamiento de unos pocos; y porque permite al usuario poder usar este sistema o no, ya que lo planteaban como una posibilidad.

Pero esto no era todo. No bastaba con buenas intenciones. En AlegrIA sabían que quedaba más por hacer.

### *Acto II: Apliquemos los principios*

> Indaga qué principios debemos incorporar para alcanzar ese fin específico, dentro de una moral cívica de la sociedad en la que se inscribe.

En la cuestión de los principios el tema estaba más claro. Sabían, gracias a aquel becario que despidieron, joven formado en temas éticos de la inteligencia artificial, que la Comisión Europea había publicado una serie de Directrices Éticas para una Inteligencia Artificial Fiable, donde aparecían los principios éticos que debían guiar la inteligencia artificial:

- Respeto de la autonomía humana. Mantener nuestra capacidad de decidir y de elegir; la inteligencia artificial no nos puede subordinar, coaccionar, engañar, manipular, condicionar o dirigir.
- Prevención del daño. La inteligencia artificial debe ser segura y robusta para que no dañe nuestra dignidad humana ni nuestra integridad física o mental.

- Equidad. Garantizar un coste y beneficio bien distribuido; evitar sesgos injustos o cualquier tipo de discriminación; poder oponernos a las decisiones que ofrezca el sistema inteligente.
- Explicación. Se tiene que poder explicar cómo funciona esta y cuál es su finalidad.

Estos principios encajaban en la sociedad en la que AlegrIA ofrecía sus servicios. Una sociedad basada en ciudadanos libres y soberanos, protegidos por criterios democráticos, bajo el gobierno de la ley, y con valores sociales sustentados en ideas de tolerancia, justicia y solidaridad. No había que buscar más principios. Pero sí había que hacer más.

Tenían que evitar que estos principios se quedaran en solo bonitas palabras. Para ello, tomaron estos principios y explicaron a sus clientes y a la sociedad cómo eran interpretados en AlegrIA. A qué se obligaban en AlegrIA para que su inteligencia artificial y su *chatbot* cumplieran con esos principios (dicho en lenguaje de Kant, cuáles eran sus imperativos categóricos):

- Respeto a la autonomía humana: el usuario sabe que empleamos una inteligencia artificial y podrá anular el uso de esta tecnología para la selección de sus contenidos.
- Prevención del daño: avisaremos al usuario si prevemos un uso excesivo que le pueda causar adicción, dándole herramientas para remediarlo.
- Equidad: revisamos continuamente nuestras fuentes de datos para evitar sesgos injustos. Utilizamos distintos criterios de evaluación de contenidos.
- Explicación: cuando te ofrezcamos una recomendación, te diremos en qué nos hemos basado.

En el Comité de Dirección empezaban a estar contentos. Hasta pensaron en volver a contratar al animoso becario, pero eso lo dejaron para más tarde.

Ahora tocaba poner en práctica estos cuidados principios. Para ello, las mismas Directrices Éticas para una Inteligencia Artificial Fiable de la Comisión Europea daba ciertas claves. Vieron que en estas directrices se hablaban de lo que llamaba requisitos clave, o algo así, y que permitían poner en práctica estos principios. Se referían a temas para garantizar la intervención de las personas, la seguridad, la privacidad, la transparencia y el bienestar social. Lo veían bien en el Comité de Dirección, pero no sabían muy bien qué hacer con aquello.

Resolvieron entonces implicar a los técnicos de AlegrIA, a sus ingenieros e ingenieras, para poner estos requisitos en práctica. Estos vieron los puntos con atención y determinaron que podían establecer acciones para que el algoritmo que ellos utilizaban cumpliera con ellos. Pero poco podían hacer con el *chatbot*, ya que éste era propiedad de la empresa Dígamebot. Había que implicarles también a ellos.

Así hicieron y explicaron su visión ética a Dígamebot. Estos no se negaron, «sí, la ética es muy importante», dijo su director de operaciones, pero pusieron varias pegas: que si la propiedad intelectual, que eso de abrirles el algoritmo…, que tenían otros clientes, que aquello les frenaba la innovación, que no siempre podían explicar el resultado, pues ya se sabe cómo es la inteligencia artificial, «y no digamos las redes neuronales», apostilló un desarrollador. Que ellos, en Dígamebot, claro que defendían la ética, pero que una cosa era una cosa y otra aquello… (gran argumento, muy habitual hacia el final de toda discusión).

—Pues veremos otras opciones… —dijeron en AlegrIA.

¡Oye!, se acabaron las reticencias. Dígamebot comenzó a trabajar con AlegrIA para definir acciones concretas en el diseño del algoritmo y del chatbot, que garantizara los principios éticos a los cuales se habían comprometido. Por ejemplo, incluyeron una opción que permitía que el usuario pudiera seleccionar qué parámetros quería que se usaran para le propusieran contenidos, o una especie de «botón de apagado» que anulaba el uso de la inteligencia artificial para su caso.

Aquello ofreció confianza y el Comité de Dirección de AlegrIA decidió publicar en sus redes sociales los principios a los que se comprometían y cómo iban a conseguirlo.

—Ahora, ¿qué toca? —preguntó Ariadna.

### *Acto III: Tenemos que explicarlo*

Conoce la situación y los medios por los que se alcanza ese fin específico, mediante información precisa y clara de cómo está actuando la inteligencia artificial.

Había que explicar todo esto: cómo funcionaba el algoritmo de AlegrIA y el *chatbot* de Dígamebot. Los técnicos de ambas empresas propusieron publicar un *paper*.

—¿Qué es eso? —preguntaron en el Comité de Dirección.

—Un artículo de estilo académico, donde se explican los parámetros técnicos de los algoritmos.

La idea no tuvo éxito. Sin menoscabo de los clientes de AlegrIA, su nivel de comprensión técnica no estaba necesariamente en niveles académicos. Hacía falta algo más sencillo, orientado al público en general.

Con gran esfuerzo por parte de los técnicos, para evitar esas palabras tan raras que suelen utilizar, se llegó a unos párrafos breves donde se contaba lo básico para entender qué es lo que hacían en AlegrIA con la inteligencia artificial. Explicaban qué fuentes de datos usaban, cuáles eran los criterios por los cuales los algoritmos ofrecían recomendaciones y qué cosas hacían y qué no hacían con la información recopilada.

Se lo presentaron al Comité de Dirección, y les gustó, porque lo entendieron.

—¿Y los riesgos? —dijo Habana, miembro del Comité de Dirección.

—¡Los riesgos! ¡Vamos a contar los riesgos! —exclamó un técnico.

—A mí me parece bien —respondió Kanza. Si creemos en unos principios, debemos hacer por cumplirlos. Uno de nuestros principios es la autonomía humana, que significa que el usuario pueda elegir. Para que pueda elegir bien, le tenemos que explicar los beneficios y posibles riesgos.

—Sí, pero… tengamos en cuenta las consecuencias… Podríamos transmitir alarmismo —intervino Benito.

—Busquemos ese punto medio —concluyó Ariadna.

Así hicieron. Además de explicar de forma sencilla y clara qué hacían con su inteligencia artificial, exponían posibles riesgos, sin el propósito de causar inquietud o desasosiego, y ofrecían a sus clientes posibles soluciones. Por ejemplo, existía un riesgo de adicción. Su sistema era tan bueno, que los contenidos ofrecidos iban a encantar a los usuarios y estos podrían pasarse horas y horas ante la pantalla viendo vídeos o películas. Para evitar este exceso, se informaría al usuario de tiempos abusivos de consumo e incluso le darían la opción de permitir que AlegrIA le cerrara temporalmente la plataforma, si así lo deseaba. También había riesgo de que AlegrIA llegara saber demasiado de sus usuarios. Para ello, explicaban lo que nunca preguntarían, lo que nunca analizarían y daban recomendaciones para que los usuarios no revelaran excesiva información.

—Esto da confianza —comentó Nieves.

—Bueno, pues ya está. Ahora les toca a nuestros clientes —dijo satisfecha Ariadna.

### *Corolario: Eres libre de decidir*

Deja la toma de decisión en manos de los afectados por la inteligencia artificial, los cuales, con esa información precisa y clara, puedan ponderar las consecuencias, sirviéndose de criterios tomados de distintas éticas.

Pues sí, la dirección de AlegrIA había hecho todo lo que estaba en su mano: habían establecido por qué usaba la inteligencia artificial y por qué era legítimo usarla; habían determinado cómo poner en práctica unos principios éticos; y habían explicado lo que hacían y no hacían con la inteligencia artificial. Ya no podían, ni debían hacer más, pues el siguiente paso quedaba en la libertad de sus clientes.

Cada usuario tenía la opción de decidir qué contenidos querían que le propusieran y de qué forma, acorde a sus valores éticos. Aquello ya era una labor personal. Cada uno debía considerar sus principios éticos, ponderar las consecuencias de hacer una cosa u otro y decidir finalmente cómo quería ser tratado por la inteligencia artificial de AlegrIA. Todo ello según ese ideal razonado de lo que juzgara digno de ser. Obviamente, AlegrIA nunca dijo esto de esta forma a sus clientes. Nunca les dijo que tuvieran en cuenta sus

principios, las consecuencias y sus virtudes. Entendían que aquello excedía sus competencias y era inmiscuirse en la forma de decidir de cada uno. Todo lo que comunicaron en su plataforma fue: «eres libre de decidir, según tus valores».

Hubo clientes que no entendieron nada. Se saltaban las breves líneas que explicaban la visión ética de AlegrIA e iban directamente a ver capítulos y capítulos de la famosa serie *El Juego del Bogavante*. Era su decisión. Pero en AlegrIA vieron que otros muchos usuarios sí consideraban los grados de libertad que la plataforma les ofrecía, y lo agradecían. Hubo un éxito incipiente que corrió por los medios de comunicación y que les hizo crecer en clientes.

—Esto de ser ético va a ser rentable —comentó en televisión un experto en inversiones tecnológicas.

Otros analistas argumentaron que AlegrIA habría tenido un crecimiento de doble dígito, «si no se hubiera impuestos tantas restricciones éticas excesivas», en palabras suyas, aduciendo que «distintos estudios vaticinaban que el uso de la inteligencia artificial en el tratamiento de datos iba a generar un crecimiento sostenido entre el 20 y el 30 % en los próximos años».

En una rueda de prensa, le preguntaron a Ariadna, por ese posible crecimiento perdido. Ella contestó que nunca se pierde lo que no se tiene y que en AlegrIA, importaba tanto lo que se conseguía, como la forma en que se conseguía.

—No queremos que nuestros clientes se sientan como una sopa de datos —concluyó la directora ejecutiva.

No todos lo entendieron.

## Evitar la inteligencia artificial pop

Latas de sopa de usuario, esa es la pintura de la inteligencia artificial pop. Esa que te convierte en una sopa datos. Si apelamos a la visión de responsabilidad sin culpa que hemos comentado, la pregunta adecuada para evitar esta sopa de datos no es ¿quién es el culpable?, sino ¿qué podemos hacer para evitarlo? Hay dos respuestas, porque hay dos actores principales: las organizaciones y los ciudadanos.

¿Qué pueden hacer las organizaciones?

La ética aplicada puede ser una solución a ese cuadro cubista de éticas. No tanto como una nueva teoría ética, sino como un marco de reflexión para dar una respuesta moral a la inteligencia artificial. Quizás no sea perfecta, y no solucione todos los problemas y dilemas éticos que nos podamos encontrar. Pero, como ya vimos, ninguna teoría ética lo es, ni lo será. Si ser éticos fuera algo matemático, ya habríamos dado con la fórmula. Y en ese caso, la vida dejaría de ser interesante.

La ética aplicada quizás no resuelva todos los casos, pero sí permite abordar bastantes. Ofrece una serie de pasos concretos, que van y vienen entre los principios y las consecuencias. Esto nos permite llevar a la práctica esas grandes palabras éticas que forman los principios, y evitar que todo se reduzca a ver qué se consigue, según una visión de las consecuencias.

Utilizar la ética aplicada por parte de las organizaciones requiere de dos actos de sinceridad. Uno al principio y otro al final.

Al principio, lo primero que pueden, y deben, hacer las organizaciones es creer firmemente en aquellas restricciones éticas que se impongan. Esa será su primera expresión de responsabilidad. Imponerse normas a uno mismo es un acto de libertad. Porque somos libres, nos imponemos leyes. Por ello, el primer paso de una organización que quiera ser ética con la inteligencia artificial es que asuma sinceramente su responsabilidad sobre lo que ella considera que está bien y está mal. La base de ser ético no es acertar siempre en tus actos, entonces, nadie sería ético, sino responder de ellos, ante ti y ante los demás.

Querer sinceramente responder. Esa es la primera sinceridad.

Una vez asumida esa responsabilidad sincera, el resto de acciones vienen determinadas por los propios pasos de la ética aplicada. Hasta llegar al último, el cual requiere el segundo acto de sinceridad: sinceramente querer que el ciudadano, en la figura de usuario de una inteligencia artificial, pueda tomar sus decisiones.

No digo con esto que el usuario pueda hacer cualquier cosa, y que no tenga límites. Eso podría llevar también a consecuencias poco éticas y desastrosas. Hablo de tener la intención sincera de no considerar a las personas como una sopa de datos.

Estos son los dos actos de sinceridad necesarios, sin los cuales los pasos de la ética aplicada se quedan en un mero procedimiento de cara a la galería. Cosa posible y tentadora en este mundo de postureo. Ese riesgo siempre existirá. Estas recomendaciones están pensadas con cariño para aquellas organizaciones que con sinceridad quieran ser éticamente responsables con la inteligencia artificial.

¿Qué podemos hacer nosotros como ciudadanos?

El último paso de la ética aplicada queda en manos de los afectados, que somos nosotros. De una manera particular, nosotros, como usuarios de una inteligencia artificial, o bien de manera general, como ciudadanos. El último paso nos cede el derecho a decidir cómo queremos obrar. No podría ser de otra forma. Ya hemos visto que no tiene sentido hablar de una inteligencia artificial ética. Nosotros somos los que debemos ser éticos con la inteligencia artificial.

En esto de buscar una inteligencia artificial ética, las organizaciones toman sus decisiones en los tres primeros pasos de la ética aplicada que hemos visto. Son los pasos que hablan del diseño de una inteligencia artificial para un cierto uso. El paso final es nuestro turno de decidir. ¿Sobre qué decidimos? Dado que somos, entre otras cosas, eso que llaman usuarios, nos toca decidir sobre ese uso de la inteligencia artificial.

Para ello, lo primero que tiene que suceder es que en el diseño de la inteligencia artificial nos hayan dejado distintas formas de uso, que dispongamos de ciertos grados de libertad. Si resulta que, por ejemplo, una aplicación con inteligencia artificial ha sido diseñada para que todo lo haga la aplicación, no habrá entonces nada que decidir. Todo lo decide la aplicación, es decir, la organización que diseñó la aplicación. Es entonces cuando pasamos de ser usuarios a ser usados. De ahí la importancia de este cuarto paso, que consiste en dejar la decisión en manos de los afectados, y de ahí la importancia de esa segunda sinceridad que antes comentaba de realmente querer dejar esos grados de libertad en los usuarios.

Esta decisión de cómo nosotros vamos a usar una inteligencia artificial dependerá de nuestros principios éticos, de la evaluación que hagamos de las posibles consecuencias y de nuestras virtudes. No puedo entrar en ello. Espero que este libro te haya ayudado a encontrar una visión propia. Por mi parte, al final del libro te daré mi visión personal, tan solo con el afán de compartirla contigo, dado que has llegado hasta aquí.

Pero, al margen de este obrar personal, sí que podemos hacer acciones conjuntas, como usuarios y como ciudadanos, que lleven a las organizaciones a un diseño ético con la inteligencia artificial, que luego nos permita a nosotros decidir. Aquí dejo algunas propuestas:

1. Lee sobre temas de inteligencia artificial, para saber cómo funciona y

qué es capaz de hacer. La inteligencia artificial es la herramienta del futuro y es bueno que sepamos cómo trabajar con ella.
2. Lee sobre temas de ética, será más complicado engañarte.
3. Busca información sobre cómo una organización está usando la inteligencia artificial en los servicios que te ofrece. Esta información debería estar publicada por la propia organización. Si no es así, corresponde investigar en fuentes solventes.
4. Investiga por los posibles riesgos de usar una inteligencia artificial, sin alarma, con la idea de buscar solución.
5. Pide grados de libertad en una aplicación, para poder decidir cómo usar la inteligencia artificial.
6. Si te dan esa autonomía, úsala; huye de ser un autómata. Si no te la dan, huye de la aplicación.
7. Lee las condiciones legales y las cookies. Son un rollo, lo sé, porque quieres entrar cuanto antes en la aplicación, pero quizás encuentres todo lo anterior.
8. Exige que haya certificaciones éticas, como existen certificaciones en otros ámbitos. Esto evita, o complica, el postureo ético.
9. Habla de ética, discute de ética, pero de forma razonada. Busca convencer, más que vencer; sabiendo que no convencer tampoco es perder.

Parecen mandamientos, pero son recomendaciones para sopesar no ser una sopa de datos. Para que uses la inteligencia artificial sin que ella te use a ti.

**Mis conclusiones. ¿Las tuyas?**

En este capítulo hemos abordado las siguientes cuestiones: ¿Qué pueden hacer las organizaciones? ¿Qué podemos hacer nosotros como ciudadanos? Estas son mis propuestas de respuesta:

- La ética aplicada es un marco que permite reflexionar sobre una actividad de nuestra sociedad desde una perspectiva moral.
- La ética aplicada busca un equilibrio entre obrar solo según unos principios (éticas deontológicas) y obrar solo según las consecuencias (éticas teleológicas). Este equilibrio se consigue con la mediación de las virtudes.
- La virtud es aquel tipo de comportamiento que da sentido a nuestra existencia, al hacer solo aquello que juzgamos digno de lo que somos o queremos ser.
- Distintas organizaciones han pensado en posibles principios éticos para la inteligencia artificial. La Comisión Europea ha definido los siguientes:

**Comisión Europea**

| Principios éticos… | …que se consiguen mediante |
|---|---|
| Respeto de la autonomía humana. | |
| Prevención del daño. | |
| Equidad. | |
| Explicación. | |
| Diversidad, no discriminación y equidad. | |
| Bienestar social y ambiental. | |
| Rendición de cuentas. | |

Puede existir un compromiso entre principios, creando un dilema. Para abordar el dilema es bueno seguir una serie de pautas:

- No crear una jerarquía de principios.
- No debatir sobre los principios de forma teórica.
- Resolver el dilema mediante un proceso de reflexión razonado.

Para evitar caer en la llamada *responsabilidad sin sujeto*, donde nadie es responsable, debemos pensar en la *responsabilidad sin culpa*. Eso significa plantear la pregunta ¿qué podemos hacer para que no vuelva a ocurrir?, en lugar de plantear ¿quién fue el culpable?

La ética aplicada consta de cuatro pasos que se pueden usar de forma práctica:

- Determinar claramente el *fin específico* que *legitima* el uso de la inteligencia artificial en una actividad.
- Indagar los *principios* dentro la *moral cívica de la sociedad* en la que se vives.
- Conocer de forma *clara y precisa* cómo está actuando la inteligencia artificial.
- Dejar que tomes una *decisión*, ponderando las *consecuencias* que la inteligencia artificial puede tener para ti, considerando *tus criterios éticos*.

Para llevar a cabo estos pasos, las organizaciones deben querer dejar decidir, y nosotros debemos querer decidir.

Estas son mis conclusiones. Pero evita ser una sopa de datos y haz tu propio caldo.

# Los robots no harán yoga

Este libro inicialmente se iba a llamar *Los robots no harán yoga*. Con ello quería destacar que la inteligencia artificial nada puede hacer con la ética. Para decidir en ese equilibrio entre los principios y las consecuencias necesitamos de las virtudes, las cuales solo se consiguen siendo uno consciente de lo que es y de lo que quiere ser. La inteligencia artificial podrá ser inteligente, pero no virtuosa, pues ni sabe quién es, ni quién quiere ser.

¿Te imaginas un robot practicando yoga? ¿Sentado en flor de loto? Posiblemente lo podría hacer, en el sentido de adoptar físicamente esa y otras posturas yóguicas. Pero, ¿tendría sentido? ¿podría así un robot practicar la consciencia?

Hablando de esto con mi buen amigo Iván, me defendió que sí, que los robots podrían llegar a hacer yoga. En ese momento no me dijo más, pero al día siguiente recibí un extenso correo electrónico suyo donde me argumentaba su respuesta. Incluso me lanzaba una nueva cuestión: ¿Qué hará el ser humano cuando los robots hagan yoga?

Su propuesta se fundamentaba en la visión de una inteligencia artificial fuerte, la cual defiende que una inteligencia artificial suficientemente compleja llegará a tener nuestro mismo nivel de inteligencia y capacidades. Con el tiempo, la inteligencia artificial podrá ser mejor científica que nosotros, mejor matemática, mejor abogada, mejor poeta, o mejor artista.

En ese proceso de igualdad con nosotros, la inteligencia artificial también llegaría a tener emociones. Incluso quizás llegara a tener procesos químicos causantes de dichas emociones. Sería entonces cuando necesitaría hacer yoga, para alcanzar un equilibrio entre sus componentes racional y emocional. Similar a nuestra necesidad, cuando hacemos yoga.

Para entonces, una vez que las máquinas fueran capaces de dar respuesta física, cognitiva y emocional igual o mejor que nosotros, ¿a qué nos dedicaríamos? Me vaticinaba, concluyendo su misiva electrónica, que «a responder las necesidades espirituales de la humanidad, hasta que, si fuera posible, pudiéramos programar un alma en las máquinas...».

Este libro discute y defiende justo lo contrario. Mi aprecio y respeto por mi buen amigo Iván me animaron a incluir su punto de vista. Porque no hay sabiduría sin contraste de puntos de vista distintos.

Según el budismo, el yoga sirve para alcanzar el camino medio, que Aristóteles llamaba la virtud. En este libro hemos hablado de virtud como ese esfuerzo razonado por hacer aquello que juzgamos digno de lo que somos o queremos ser. Es un esfuerzo personal, que tiene que partir de la reflexión y la consciencia de cada uno. Como dije en su momento, poco puedo hacer aquí. Pero con el ánimo de ofrecer un comienzo por algún sitio, te propongo lo que llamo virtudes tecnológicas, que es mi visión personal de una cierta virtud en nuestra relación con la tecnología y la inteligencia artificial.

Las virtudes tecnológicas comprenden tres actitudes: ser más alegres, más amables y más artistas.

Ser alegre no es tanto ser feliz. La felicidad es la sonrisa pasajera en Instagram. Estar alegre no es necesariamente andar por la vida con los dedos en V. Es vivir más la vida que tenemos, es ensanchar el alma, sin estar sujeto a un placer concreto. Y es vivir junto con otros.

Un vivir con otros que nos lleva a ser amables. Amables en el sentido tan natural, sencillo y tierno de poder ser amados, de que otro te pueda apreciar por lo que haces. Esto solo es posible por nuestra naturaleza libre. Si somos amables, es decir, si mostramos afecto, cariño o bondad, es porque queremos ser amables, no porque debamos ser amables. Si nuestra naturaleza nos condicionara a mostrar afecto todo el rato, eso no sería dingo de mérito; sería consecuencia de nuestra naturaleza. ¿Es amable Siri, Alexa o Google Assistant cuando te saluda al llegar a casa, o es consecuencia de cómo está programada? Por eso la gente te aprecia cuando eres amable, porque podrías no serlo.

Esa libertad, ese querer ser o hacer algo, es la semilla de un vivir personal, un vivir único, que es la cualidad de todo artista. No hablo de tener la habilidad de pintar un cuadro o escribir una sinfonía; me refiero a ser uno mismo en todo lo que haces. Huir de una actitud copia-pega de comportamiento y forma de pensar, acomodándose a lo que hace la gente, porque eso es lo normal. Es buscar esa identidad que relata tu existencia de forma única. ¡Esa es la esencia del artista!

Virtudes tecnológicas que se pueden aplicar a cualquier dimensión de nuestra vida, y también a la inteligencia artificial. Si una inteligencia artificial te propone algo que choca con tus principios, puedes pensar si las consecuencias de aquello que vas a hacer te llevan a ser más alegre, más amable o más artista. Es solo mi visión y mi propuesta.

Este libro finalmente se ha llamado *Estupidez artificial*, con el ánimo de buscar una sabiduría natural. ¿Qué sabiduría? Podría dar para otro libro, pero quizás lo explique mejor este breve cuento[115].

Después de años de entrenamiento, el discípulo pidió al maestro que le

otorgara la sabiduría. El maestro le condujo a un bosquecillo de bambúes y le dijo: «Observa qué alto es ese bambú. Y mira aquel otro, qué corto es». Y en aquel mismo momento el discípulo entendió qué era la sabiduría.

# Fuentes, por si quieres seguir bebiendo

Todo lo que te he contado está refrendado por fuentes solventes. No obstante, harás bien en dudar de mi palabra y verificar cuanto he dicho. Aquí tienes mis fuentes, para que revises mis afirmaciones y, si te interesa, que puedas seguir bebiendo y creando.

# Agradecimientos

Cuando ya llevaba avanzado este libro surgió la gran pregunta de todo escritor: ¿cómo lo puedo publicar? Como ves, pude responder esa pregunta. Fue gracias a Jeanne Bracken, a quien agradezco su atención conmigo, por su orientación y respuestas a tantas preguntas que tenía sobre este complicado mundo de la edición de libros. Como agradezco el desprendimiento de mis mecenas, que han apoyado ilusionados mi ilusión.

# Mecenas

# We the Humans

**A**
Adolfo San Martín
Adolfo Torres
Alberto Iglesias Fraga
Alberto Mayo
Alejandra Delgado
Alejandro Jiménez
Álvaro Bernad
Ángel Crespo
Ángela Robles
Anónimo
Antonio Crespo
Antonio Cruz
Antonio José Rodríguez Méndez
Antonio Medina Díaz
Antonio Toscano
Aurora

**B**
Baltasar Carretero
Beatriz Blanco Pérez
Beatriz Hierro Alonso
Blanca CL

**C**
Carlos Simón Gallego

Carolina Borras
César Elvira
Cristina Mingo

**D**
Dani Carrión
Daniel Arquero
Daniel De la Rica Feduchi
Daniel Osorio
Daniel Santiago
David Hernández Sanz
Diana R. Cobaleda
DigitalBCN
Domingo Gaitero

**E**
Eduardo Martín
Eduardo Rodríguez
Emilio Tovar
Emilio Vianco
Erik Fernández Santos
Eugenio Criado Carbonell

**F**
Felipe Franco de Rueda
Feliz Belso Corral
Fernando García Pascua
Fernando R. Arroyo
Francisco José Maregil Nieto

**G**
Gonzalo
Gooder

**H**
Helena Cebrián

**I**
Ignacio G.R. Gavilán
Ignasi Prado Julià
Íñigo Sarría Martínez de Mendibil
Iván Martín Bermejo

**J**

Jaime Izquierdo Pereira
Javier Alonso
Jesús Gómez
Jesús Plaza Rubio
José Antonio Zamborain
José Carlos González Cristóbal
José Luis Fanjul
José Luis Martínez
José Luis Mora
José Miguel Mohedano Martínez
José Miguel Palou Larrañaga
JrouyetC
Juan Antonio Zafra
Juan Camarillo Casado
Juanjo Cukier
Juanma González

**L**
Laura García Mozo
Leopoldo
Luis Miguel Quintas Martínez
Luis Muñoz

**M**
M. ª Carmen Romero Ternero
M. ª Carmen
Maciej Krajewski
Marcos Navarro
María Asunción Gil
María González
Mariano Ferrera
Marisa Nieto-Márquez
Miguel Ángel Fernández
Miguel Díaz-Pache Gosende

**N**
Néstor Rouyet Ruiz

**O**
Odiseo
Oscar Carbajo

**P**

Pablo Ayuso
Paloma Casado
Paula Sánchez Ruiz
Pedro Irujo
Penélope
Pilar Castellano

**R**
Raquel García León
Raúl Belber
Remo Tamayo
Reyes Casado
Ricardo Antuña García
Ricardo Sinde Ceniza
Rubén M. ª Carmen González Crespo

**S**
Stéphanie Novais

**V**
Virginia Durán

**W**
We The Humans

**Z**
Zoltan A. Sánchez

Imagina un mundo sin imprenta, sin coches, sin teléfonos móviles... ¿Cuesta pensarlo, verdad? Pero si algo tienen en común estas revoluciones es que las tres tajeron consigo un fuerte cambio de mentalidad, sacudiendo los cimientos de la sociedad del momento. Algo similar estamos viviendo en la actualidad con la inteligencia artificial (IA). *Estupidez artificial. Cómo usar la inteligencia artificial sin que ella te utilice a ti* tiene el objetivo de hacerte reflexionar, desde un punto de vista filosófico, sobre el miedo infudado que se le tiene; a la vez que te invita pensar en todas sus ventajas prácticas, realizando un alegato a su uso ético y responsable y sin miedos. ¿Es cierto que la IA decide por nosotros? ¿Tenemos que creer en lo que dice la IA como si fuera la sabiduría máxima? En las páginas de este libro encontrarás la respuesta a estos y otros interrogantes, o mejor dicho... podrás encontrar tu propia opinión.

**Juan Ignacio Rouyet** (Madrid 1967) es doctor en Informática e Ingeniero de Telecomunicación. Actualmente es *Senior Manager* en la consultora Eraneos, y profesor en la UNR y en la Universidad Francisco de Vitoria, donde colabora con el Centro de Estudios e Innovación en Gestión del Conocimiento (CEIEC) en una IA ética. Además es presidente de *We the Humans Think Tank* que busca soluciones éticas para la IA. Escribe en la columna «Homo Digitalis» de *Digital Biz*, y colabora con distintos medios de comunicación. Pero ante todo, Juan Ignacio es y siempre será un ingeniero humanista.

# Notas

[1] Shah, I. (1993). «El diagnostico». *La sabiduría de los idiotas. Cuentos de tradición sufí*. Madrid: Edad.
[2]Brenson, M. *How the Spiritual Infused the Abstract*. New York Times, 21 de diciembre de 1986, sección 2, página 1.
[3] Kandinsky, W (2020). *De lo spiritual en el arte*. Traducción, Genoveva Dieterich. Ediciones Paidós.
[4] Newsome, D. (1997). *The Victorian World Picture*. New Brunswick, New Jersey: Rutgers University Press. El detalle de cómo sucedieron los acontecimientos de su atropello se puede ver en: Simon Garfield (2002). *The Last Journey of William Huskisson*. London Faber and Faber.
[5] *Ibídem*, pp. 31-32.
[6] Shepardson, D. «Uber disabled emergency braking in self-driving car: U.S. agency». *Reuters*, 24 de mayo de 2018
[7] Milne-Smith, Amy. «Shattered Minds: Madmen on the Railways, 1860–80». March 2016, Journal of Victorian Culture, Vol. 21, págs. 21-39.
[8] Torrey, E. F. y Miller, Judy. (2002). *The Invisible Plague: The Rise of Mental Illness from 1750 to the Present*. New Jersey: Rutgers University Press, pp. 98-99.
[9] Godwin, G. (1837). *An appeal to the public, on the subject of railways*. London: J. Weale, J. Williams, pp. 14-16.
[10] *Ibídem*, p. 31.
[11] Jackman, W.T. (1916). *The Development of Transportation in Modern England. Vol. II*. Cambridge: Cambridge University Press, pp. 485 y ss.
[12] *Ibídem*, pp. 39-40.
[13] Jackman, W.T. (1916). *The Development of Transportation in Modern England. Vol. II*. Cambridge: Cambridge University Press, pp. 497-499.
[14] Godwin, G. (1837). *An appeal to the public, on the subject of railways*. London: J. Weale, J. Williams, p. 14.
[15] *Ibídem*, p. 41.
[16] *Ibídem*, p. 45.
[17] Czitrom, D. J. (1982). *Media and the American Mind. From Morse to McLuhan*. The University of North Carolina Press, pp. 9-10.
[18] *The Intellectual Effects of Electricity*. Spectator. November 9, 1889, págs. 11-12. Este mismo periódico en noviembre de dicho año publicó otro editorial lamentando que en breve se iba a editar un periódico con ilustraciones. Decía que obviamente dicho tipo de periódico sería para personas sin capacidad de imaginación, y que necesitaban dibujos para entender las noticias. En este editorial de nuevo arremetía contra el telégrafo y sus consecuencias en los periodistas.
[19] Casson, H. N. (1922). *The History of the Telephone*. Chicago: A. C. McClurg & Co., p. 39.
[20] *Ibídem*, p. 42-45.

[21] *Ibídem*, p. 247.
[22] Marvin, C. (1988). *When Old Technologies Were New. Thinking About Electric Communication in the Late Nineteenth Century*. New York: Oxford University Press, p. 68.
[23] *Ibídem*.
[24] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
[25] Rawlinson, K. *Microsoft's Bill Gates insists AI is a threat*. BBC News, 29 de enero de 2015.
[26] Gibbs, S. Elon Musk: artificial intelligence is our biggest existential threat. The Guardian, 27 de octubre de 2014.
[27] Cellan-Jones, R. Stephen Hawking warns artificial intelligence could end mankind. BBC News, 2 de diciembre de 2014.
[28] OECD. OECD Employment Outlook 2019. OECD.
[29] Bidadanure, J. *The Political Theory of Universal Basic Income*. Annual Review of Political Science 2019. Vol. 22, pp. 481-501.
[30] Goodfellow, I., Shlens, J., Szegedy, C. *Explaining and Harnessing Adversarial Examples*. ICRL 2015.
[31] Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M. *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*. ACM CCS'16 October 24-28, 2016, Vienna, Austria.
[32] Mill, J. S. (1976). *Sobre la libertad*. Madrid: Alianza Editorial.
[33] Ortigosa, A., Carro, R., Quiroga, J. *Predicting user personality by mining social interactions in Facebook*. Journal of Computer and System Sciences, vol. 80, nº 1, pp. 57–71.
[34] Goldberg, L.R. *An Alternative Description of Personality: The Big-Five Factor Structure*. Journal of Personality and Social Psychology, vol. 59, nº 6, pp. 1216–1229.
[35] Amichai-Hamburger, Y., Vinitzky, G. *Social Network Use and Personality*. Computers in Human Behavior, vol. 26, nº 6, pp. 1289-1295.
[36] Reese, H. *Why Microsoft's 'Tay' AI bot went wrong*. Tech Republic, 24 de marzo de 2016.
[37] Zhao, J., Wang, T., Yatskar, M., Ordonez, V. y Chang, K. *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*. Conference: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2979-2989.
[38] Coughlan, S. *Why did the A-level algorithm say no?* BBC, 14 de agosto de 2020.
[39] Cross, S. (1998). *After Mountains and Sea: Frankenthaler,1956-1959*. Stuttgart: Guggenheim Museum Publications.
[40] Turing, A. M. *Computing Machinery and Intelligence*. Mind, Vol. 49, 1950, pp. 433-460.
[41] Weizenbaum, J. *ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine*. Communications of the ACM. Vol. 9, Nº 1, 1966, pp. 36–45.
[42] Welch, C. *Google just gave a stunning demo of Assistant making an actual phone call*. The Verge, 8 de mayo de 2018.
[43] Peter, J. *Artificial versifying or The school-boy's recreation. A new way to make Latin verses. Whereby any one of ordinary capacity, that only knows the A.B.C. and can count 9 (though he understands not one word of Latin, or what a verse means) may be plainly taught, (and in as little a time as this is reading over,) how to make hundreds of hexameter verses, which shall be true Latin, true verse, and good sense. Never before publish'd*. Charleston SC: Proquest, Eebo Editions, 2011.
[44] Strachey, C. *The "Thinking" Machine*. Encounter, octubre 1954, pp. 25-31. Se puede acceder a un simulador de M.U.C. en https://www.gingerbeardman.com/loveletter/ [último acceso 18/10/2022]
[45] Lutz, T. *Stochastische Texte*. Augenblick, vol 4, Nº 1, 1959, pp. 3-9
[46] Hofstadter, D., Dennett, D. (eds.). *The Mind's I: Fantasies and Reflections on Self and Soul*. New York, Basic Books, 1981. Capítulo 26: A Conversation with Einstein's Brain.
[47] Obvious. *La familia de Belamy*, s.f. Recuperado de https://obvious-art.com/la-famille-

belamy [último acceso 18/10/2022]
[48] Fautrel, P., Caselles-Dupré, H., Vernier, G. *Obvious Manifesto*, s.f. Recuperado de https://obvious-art.com/page-about-obvious [último acceso 18/10/2022]
[49] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. *Generative Adversarial Networks*. Advances in Neural Information Processing Systems. 3 (11), 2014.
[50] Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M. *CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms*. VIII International Conference on Computational Creativity (ICCC), Atlanta, GA, 20-22 de junio de 2017.
[51] Russell, S., Norvig, P. (2010). *Artificial intelligence: a modern approach* (3.ª edición). Prentice Hall, Upper Saddle River, New Jersey, pp 610 y ss.
[52] Allais, M. (1953). *Le comportment de l'homme rationnel devant la risque: critique des postulats et axiomes de l'´ecole Am´ericaine. Econometrica*, 21, 503–546.
[53] Paquet, M. (2019). *René Magritte*. Colonia: Taschen Benedikt.
[54] Breton, A. (2001). *Manifiestos del surrealismo*. Buenos Aires: Argonauta, p. 44.
[55] Appollinaire, G. (2018). *Las tetas de Tiresias*. Kroebel. Zaragoza: Libros del Innombrable.
[56] McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August, 1955.
[57] Legg, S, Hutter, M. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*. Proceedings of the AGI Workshop 2006. Goertzel and P. Wang (Eds.), pp. 17-24. IOS Press, Amsterdam. En este artículo se pueden encontrar más de 70 definiciones sobre qué es inteligencia, y los autores aportan una más de su cosecha.
[58] Gottfredson, L. S. *Mainstream Science on Intelligence (editorial)*. Intelligence. Volume 24, Issue 1, January–February 1997, pp. 13-23.
[59] Descartes, R. *Meditaciones metafísicas*. Meditación II.
[60] Kant, E. *Prolegómenos*. Párrafo 22.
[61] Searle, J. R. *Minds, brains and programs*. Behavioral and Brain Sciences. Vol. 3, 1980, pp. 417-424.
[62] Schank, R., Abelson, R. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Hillsdale, New Jersey: Lawrence Eribaum, 1977, pp. 42-46.
[63] Turing, A. M. *Computing Machinery and Intelligence*. Mind, Vol. 49, 1950, pp. 433-460.
[64] Jefferson, G. *The Mind of Mechanical Man*. British Medical Journal. 1 (4616), 1949, pp. 1105–1110.
[65] Wiener, P. (Ed) (1951). *Leibniz: Selections*. New York: Charles Scribner's Sons, p. 51.
[66] Hilbert, D., Ackermann, W. *Grundzüge der theoretischen Logik* (Principios de la lógica matemática). Springer-Verlag, 1928.
[67] Church, A. *An Unsolvable Problem of Elementary Number Theory*. American Journal of Mathematics, 1936, vol. 58, nº 2, pp. 345-363.
[68] Turing, A. *On Computable Numbers, with an Application to the Entscheidungsproblem*. Proceedings of the London Mathematical Society, 1936, Series 2, nº 42, pp. 230–265. Errata corregida en Series 2, nº 43 (1937), pp 544–546.
[69] Penrose, R. (1989). *Emperor's New Mind*. Oxford University Press.
[70] Russell, S., Norvig, P. *Artificial intelligence: a modern approach* (3.ª edición). Prentice Hall, Upper Saddle River, New Jersey, 2010.
[71] Harnad, S. *Minds, Machines and Searle*. Journal of Theoretical and Experimental Artificial Intelligence, 1989, vol. 1, pp. 5-25.
[72] Poole, D.; Mackworth, A. y Goebel, R. (1998). *Computational intelligence. A logical approach*. New York: Oxford University Press.
[73] Platón (2018). *Diálogos*. Vol. I. Madrid: Gredos.
[74] Serrat, J. M. (1981). *Esos locos bajitos*. Álbum "En tránsito". Discográfica Ariola (Sony Music), Madrid, mayo de 1981.

[75] Cortina, A. y Martínez, E. (1996). Ética. Madrid: Akal.
[76] Camps, V. (2013). *Breve historia de la ética*. Barcelona: RBA.
[77] Aristóteles. Ética *a Nicómaco*. El contenido explicado se puede encontrar en el Libro I (capítulos 1, 2, 7, 8, 13), Libro II (capítulo 6), libro IV (donde describe las virtudes), Libro VI (capítulo 5) y Libro VIII (capítulo 10).
[78] Santo Tomás. *Summa Theologica*. El contenido explicado se puede encontrar en la parte *Prima Secundae*, en las siguientes cuestiones (c) y artículos (a): c1, a7; c3, a8; c4, a4; c5, a5, a7; c55, a4; c61, a1; c62, a1; c63, a2; c91, a2, a4; c94, a3.
[79] Hume. *Tratado de la naturaleza humana*. El contenido explicado se puede encontrar en los siguientes Libros (L), Partes (P) y Secciones (s): LII, PI, s11; LII, PIII, s3, s6, s10; LIII, PI, s1, s2; LIII, PIII, s1, s3.
[80] Kant. *Fundamentación de la metafísica de las costumbres*. El contenido explicado se encuentra repartido en los tres capítulos que conforman la obra. Las formulaciones del imperativo categórico se encuentran en el capítulo 2.
[81] Asimov, I. (1942). *Runaround*. Nueva York: Astounding Science Fiction, Street & Smith.
[82] Asimov, I. (1950). *I Robot*. Nueva York: Gnome Press.
[83] Bentham, J. *An Introduction to the Principles of Morals and Legislation*. El contenido explicado se puede encontrar en el capítulo I, párrafos I a X; capítulo IV, párrafos del I al V.
[84] Mill, J.S. *Utilitarism*. El contenido explicado se puede encontrar en el capítulo 2.
[85] Nietzsce, F. *Aurora*. Párrafo 103.
[86] Nietzsche, F. *Más allá del bien y del mal*. El contenido explicado se puede encontrar en los párrafos 13, 22, 39, 186, 187, 207, 225, 227, 228, 259, 260, 266, 272, 290.
[87] Nietzsche, F. *El caminante y su sombra*. Párrafo 52.
[88] Nietzsche, F. *Genealogía de la moral*. Segundo Tratado, párrafos 24 y 25.
[89] Nietzsche, F. *Así habló Zaratrusta*. El contenido explicado se puede encontrar en los capítulos «Prologo de Zaratrusta», «De las tres transformaciones», «El convaleciente».
[90] Habermas, J. (1985). *Conciencia moral y acción comunicativa*. Península, Barcelona. El contenido explicado se puede encontrar en las páginas 86, 88, 110 a 117 y 143.
[91] *Ibídem*, p. 76.
[92] Filipowska, R. (2016). *Richard Hamilton's Plastic Problem*. Distillations, 15 de octubre de 2016. https://www.sciencehistory.org/distillations/magazine/richard-hamiltons-plastic-problem [último acceso 18/10/2022].
[93] Comenas, G. (2022). *The Origin of Andy Warhol's Soup Cans or The Synthesis of Nothingness*. Warholstars.org. https://www.warholstars.org/andy_warhol_soup_can.html [último acceso 18/10/2022].
[94] Foot, P. (1967). *The Problem of Abortion and the Doctrine of the Double Effect*. Oxford Review, No. 5.
[95] Bonnefon, J-F., Shariff, A., Rahwan, I. (2016). *The social dilemma of autonomous vehicles*. Science, 24 Jun 2016, Vol 352, Issue 6293, pp. 1573-1576. *La plataforma Moral Machine* se puede encontrar en https://www.moralmachine.net [último acceso 18/10/2022].
[96] Awad, E., Dsouza, S., Shariff, A., Rahwan, I., Bonnefon, J-F (2020). *Universals and variations in moral decisions made in 42 countries by 70,000 participants*. Psychological and Cognitive Sciences, January 21, 2020, Vol 117, Issue 5, pp. 2332-2337.
[97] Weber, M. (2012). *El político y el científico*. Alianza Editorial, Madrid. La obra contiene el discurso "La política como vocación".
[98] Etxeberria, X. (2002). *Temas básicos de ética*. Bilbao: Desclée.
[99] HWE (1979). *Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Department of Health, Education, and Welfare, Washington, D.C.
[100] Comte-Sponville, A. (2003). *Diccionario Filosófico*. Traduccion Jordi Terré. Barcelona: Paidós.
[101] MacIntyre, A. (1987). *Tras la virtud*. Barcelona: Crítica, p. 271.

[102] Camps, V. (2013). *Breve historia de la ética*. Barcelona: RBA, p. 399.
[103] Cortina, A. y Martínez, E. (1996). *Ética*. Madrid: Akal.
[104] Camps, V. (2013). *Breve historia de la ética*. Barcelona: RBA, pp. 402-409.
[105] HWE (1979). *Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Department of Health, Education, and Welfare, Washington, D.C.
[106] Puedes ver, por ejemplo: Murphy, R. y Woods, D. (2009). *Beyond Asimov: The Three Laws of Responsible Robotics*. Intelligent Systems, IEEE 24(4), pp. 4-20; Winfield, A. (2011). *Five roboethical principles – for humans*. New Scientist, 4th May.
[107] ACM (2017). *Statement on Algorithmic Transparency and Accountability*. Association for Computing Machinery US Public Policy Council (USACM), 12 january 2017.
[108] Burt, A. (2019). *The AI Transparency Paradox*. Harvard Business Review, 13 december 2019.
[109] FAT/ML (2018). *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*. Página web de FAT/ML: https://www.fatml.org/resources/principles-for-accountable-algorithms [último acceso 18/10/2022].
[110] UE (2019). *Ethics Guidelines for Trustworthy AI*. Eurpean Commission, Bruselas.
[111] Gracia, D. (2012). *Ética profesional y ética institucional: entre la colaboración y el conflicto, en La ética de las instituciones sanitarias: entra la lógica asistencial y la lógica gerencial*. Cuadernos de la Fundació Victor Grifols i Lucas, Barcelona, pág. 20.
[112] Camps, V. (2019). *Virtudes públicas. Por una ética pública, optimista y feminista*. Barcelona: Arpa, pp. 95-96.
[113] *Ibídem*, p. 100.
[114] Universidad de Cambridge. *COVID-19 Sounds App*. Los objetivos indicados se pueden ver en: https://www.covid-19-sounds.org/es/ [último acceso 18/10/2022].
[115] De Mello, A. (1989). «Los bambúes». *El canto del pájaro*. Santander: Sal Terrae.

# Otros títulos publicados

**Javier López López**
El cántico de Cthulhu

**Álvaro D. María**
Micrópolis. Más allá del Leviatán

**Jorge Segura Romano**
Céntimos underground